

Conservatoire national des arts et métiers
Architectures des systèmes informatiques
ANNEXE 03
Architecture RISC, exemple du PowerPC
Année 2002-2003

1. INTRODUCTION

Cette note présente un processeur Risc. Elle peut être lue sans référence au cours proprement dit. Toutefois, le lecteur pourra se reporter au [chapitre 3](#) des notes de cours pour les questions relatives à l'anticipation, au recouvrement et au pipeline, et au § 2.2.1 pour une [présentation comparée](#) des machines Risc et Cisc.

En matière de microprocesseurs et micro ordinateurs, la situation en 1990 était sensiblement la suivante :

- une domination très nette d'Intel, son 80486 était livré à 33MHz et le Pentium en cours d'étude;
- un effritement des positions de la série 68000 de Motorola. Son débouché en microordinateurs est limité aux machines d'Apple qui ne conquièrent plus de parts de marché. Le 68040 ne sera livré qu'en janvier 1991 avec un an de retard sur les premières annonces et l'annonce du processeur Risc 88000 n'aura pas de suite;
- IBM dont le bus MCA de 1988 n'a pas le succès attendu est de plus en plus lié à Intel et subit toujours la concurrence des clones.

Motorola, IBM et Apple décident de reconquérir des parts de marché en introduisant pour la première fois des processeurs Risc dans les machines à très grande diffusion. Leurs objectifs sont :

- de concurrencer Intel mais hors son propre terrain qui est le CISC;
- de produire rapidement des processeurs beaucoup plus puissants que ceux d'Intel;
- de pouvoir ainsi émuler ou simuler les machines du concurrent et récupérer la base installée des logiciels x86.

Pour cela, ils créent en septembre 1991 un consortium avec cinq objectifs :

- le *développement* de microprocesseurs;
- la *technique à objets* concrétisée par la création de Taligent par Apple et IBM. Cette société doit écrire un système d'exploitation à objets nommé OOS pour «object oriented operating system»;
- le *multimédia* concrétisé par la création de Kaleida par Apple et IBM. Cette société doit créer des produits pour le grand public;
- l'*interconnexion* et les réseaux;
- un *environnement de système ouvert* «poweropen environment» du genre unix pour les Macintosh et Aix.

Les microprocesseurs qui résultent du premier objectif ont une architecture commune nommée PowerPC. Elle hérite nombre de ses concepts du Power «performance optimized with enhanced risc» d'IBM. On a dit aussi «power performance chip» ou PPC. Le premier processeur, nommé *PowerPC 601*, est disponible en avril 1993 après un délai de fabrication très court.

L'architecture PowerPC est analysée dans la suite du document sur l'exemple du *PPC601* dans son mode 32 bits. Les différences entre les modes 32 bits et 64 bits sont mineures.

2. Une famille de processeurs

Le PowerPC 601 est le premier de la famille. Il est compatible avec ses prédécesseurs de la famille Power d'IBM jusqu'aux instructions et aux registres. Les nouveaux systèmes seront ainsi utilisables sans recompilation des applications existantes. Les cibles commerciales sont les matériels embarqués, les cartes additionnelles, les ordinateurs personnels, les serveurs (RS/6000) et les machines multiprocesseurs.

Le PowerPC 603 est commercialisé en octobre 1993. Il est plus élaboré et consomme moins. Ses cibles commerciales sont les portables, les matériels embarqués et les cartes additionnelles. Le fonctionnement en multiprocesseur n'est pas normalement envisagé.

Le PowerPC 604 est commercialisé au printemps 1995, sa sortie a été retardée pour cause de bogue [J1]. Son objectif commercial va des stations de travail aux machines multiprocesseur en passant par les stations graphiques ou multimédia d'entrée de gamme et les serveurs de puissance moyenne.

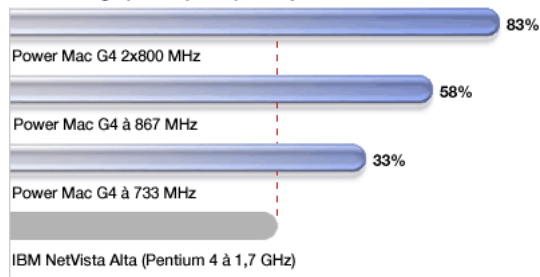
Le PowerPC 620 est le processeur 64 bits de la famille. Il est donné pour plus puissant que le PentiumPro mais ses performances le classent comme un des processeurs RISC les moins puissants de sa catégorie.

Le G4 annoncé en 1998 est disponible à plus de 500 MHz (733, 800 et 867 MHz). Il utilise une technique cuivre de 0,18 µm. Il a selon les configurations deux ou trois caches, L1 de 32 ko et L2 de 2 Mo au plus intégrés au processeur et L3 de 2Mo externe. Il est alimenté par un «MaxBus» de 128 bits. Le fonctionnement peut-être bi processeur. Les G4 ont commercialisés en 2000.

Un point intéressant. Il lui a été ajouté un dispositif dit Altivec ou «velocity engine». Cet appareillage est une extension du MMX des concurrents. Alors que ce dernier utilisait des registres de l'unité de virgule flottante pour faire des calculs entiers, on a ici une unité autonome à registres de 128 bits opérant sur deux nombres flottants simultanément ou plusieurs entiers. Il s'agit donc d'un petit mais vrai opérateur vectoriel qui apporte une spécialisation au G4 commandé pour cela par 162 nouvelles instructions. Il atteint le Gigaflops.

Apple présente ses performances sur un banc d'essai limité à certaines opérations du logiciel Photoshop comme suit.

Pourcentage plus rapide qu'un système Pentium :



Des PowerPC spécifiques existent pour l'électronique professionnelle sous la forme de microcontrôleurs.

3 ARCHITECTURE DE LA FAMILLE PowerPC PRÉSENTÉE À TRAVERS LE PPC 601

3.1. Généralités

Une architecture de microprocesseur se compose essentiellement :

- d'un modèle de mémoire tant externe, mémoire principale, qu'interne, caches;
- d'une définition des processus de traitement de programmes (ordonnancement des instructions);
- d'un jeu d'instructions pour gérer les deux précédents, il constitue la partie visible par le programmeur.

Une architecture étant définie, il peut y en avoir des réalisations diverses qui différeront par la consommation électrique, la fréquence de fonctionnement, etc. Les différences entre les architectures CISC et RISC sont visibles par leurs instructions.

3.1.1. Architecture RISC (Reduced Instructions Set Computer)

Les instructions RISC sont courtes et chacune fait une opération simple. Le nombre d'instructions est petit. D'autres caractéristiques sont importantes :

- toutes les instructions ont la même taille;
- toutes les instructions ont la même durée d'exécution;
- le nombre de formats est petit;
- la modification implicite des codes de conditions n'est pas utilisée;
- la microprogrammation n'est pas employée;
- les données sont alignées sur des adresses multiples de leur longueur;
- le nombre de registres est important;
- il faut plus d'instructions RISC que d'instructions CISC pour faire un même traitement.

Tout ceci a pour conséquences favorables :

- une compilation simplifiée par le fait que l'on emploie peu de primitives;
- des puces plus petites et de la place pour des registres;
- une consommation électrique plus faible;
- une durée de conception et des délais de commercialisation diminués.

3.1.2. Architecture CISC (Complex Instructions Set Computer)

Dans son principe, le CISC vise à réduire l'écart entre matériel et logiciel. Ses instructions de base masquent le fonctionnement du matériel. Elles sont proches des instructions nécessaires aux systèmes d'exploitation et aux langages de haut niveau. Ceci est réalisé par l'emploi systématique de la microprogrammation. Cependant, les processeurs CISC modernes s'inspirent des architectures RISC et l'exécution d'un bon nombre de leurs instructions simples est câblée.

3.1.3. Pipeline

La technique de pipeline consiste à répartir la totalité de l'exécution d'une instruction machine entre plusieurs sous unités de traitement séquentielles, appelées étages, à la manière taylorienne de la division du travail sur une chaîne. Ce travail à la chaîne est une technique utilisée dans les architectures RISC en mode synchrone comme dans les CISC en mode asynchrone. Bien souvent les performances de telle ou telle machine sont dues plus à la qualité du pipeline qu'à une vertu particulière de son architecture.

«Chaque étage du pipeline est constitué de parties combinatoires ou de mémoires à grande vitesse sous forme de caches ou de bancs de registres. Les étages sont séparés par des éléments mémorisants» [Ll p 7].

Un signal d'horloge commun donne la cadence à cette chaîne d'assemblage. L'idéal est que chaque étage exécute son travail sur une instruction en un cycle d'horloge. Cependant les dépendances entre les étages du pipeline et envers les éléments extérieurs à celui-ci (en particulier cache ou mémoire) peuvent bloquer le pipeline et diminuer ainsi son rendement.

Une bulle est le nom de ce blocage.

Annexe 3, Architecture RISC

La latence est le temps de parcours du pipeline pour une instruction.

La profondeur est le nombre d'étages du pipeline.

3.1.4. Superscalaire

Les premiers étages de lecture et de décodage qui opèrent avant que la nature de l'instruction ait été discernée sont nécessairement uniques. Ensuite, il peut y avoir un pipeline unique, c'est une architecture scalaire. S'il y a plusieurs sous pipelines spécialisés dans le traitement d'instructions de types différents, l'architecture est dite superscalaire.

Un pipeline superscalaire mène le décodage et l'exécution de plusieurs instructions dans le même temps. Il est à noter que les architectures superscalaires ou parallèles pourraient être réalisées avec des outils autres que des pipelines. Les distributions ou exécutions d'instructions en parallèle conduisent parfois à ce que les résultats sont produits dans un ordre différent de celui des instructions du programme. De plus, dans les architectures parallèles ou à base de pipelines, des accès concurrents aux ressources se produisent. La mise en œuvre de ces architectures n'est pas simple.

3.2. Unités fonctionnelles et fonctionnement général du coeur du processeur

3.2.1. Présentation générale du PowerPC 601

Deux vues d'ensemble du PPC 601 sont données dans les deux figures suivantes.

Le pipeline du PPC 601 (figure 2) constitue le coeur du processeur. Il peut être vu comme constitué de 27 étages [J2], en réalité six unités qui les regroupent. Trois de ces 6 unités sont communes à toutes les instructions :

- l'unité de chargement des instructions Fetch Unit pour FU;
- le cache unifié UGM, unité de gestion de la mémoire ou memory management unit;
- l'unité de distribution des instructions DU pour Dispatch Unit.

Les trois unités restantes, appelées unités d'exécution ou unités fonctionnelles constituent la partie superscalaire de l'architecture du processeur. Ce sont :

- l'unité de branchement BU pour «branch unit»;
- l'unité de calcul entier FXU pour «fixed point unit»;
- l'unité de calcul en virgule flottante FPU pour «floating point unit».

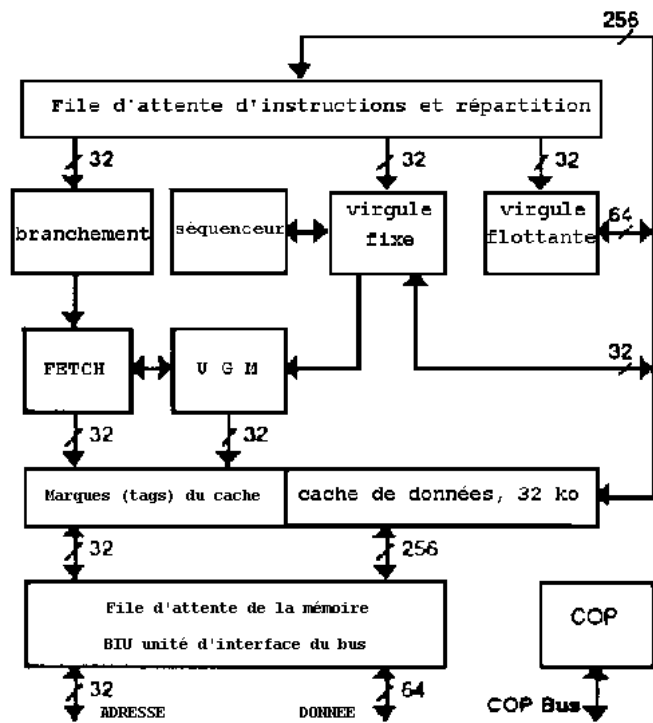
On ajoute le séquenceur utilisé par la FXU à ces 6 unités. Il gère des fonctions importantes. La FXU lui délègue plusieurs fonctions de commande du système. Ce séquenceur contient un Ko de microcode [C2], comme quoi la séparation entre CISC et RISC n'est jamais totale!

Les trois unités fonctionnelles du PPC601 sont directement calquées sur le modèle logique de l'architecture PPC. Les PPC plus récents offrent un découpage physique différent de cette logique. Chaque unité fonctionnelle correspond à un sous pipeline. Chacune possède ses propres registres et interfaces (files d'attente), y compris pour l'accès au cache. Le parallélisme entre les unités fonctionnelles est ainsi mieux géré et les problèmes de concurrence d'accès et de synchronisation mon cruciaux. Les trois unités fonctionnelles seront décrites plus loin. Chacune d'elles peut recevoir une instruction en un même cycle d'horloge. Elles constituent les derniers étages du pipeline du PPC 601. Les résultats sont stockés dans des registres spécifiques à chaque unité.

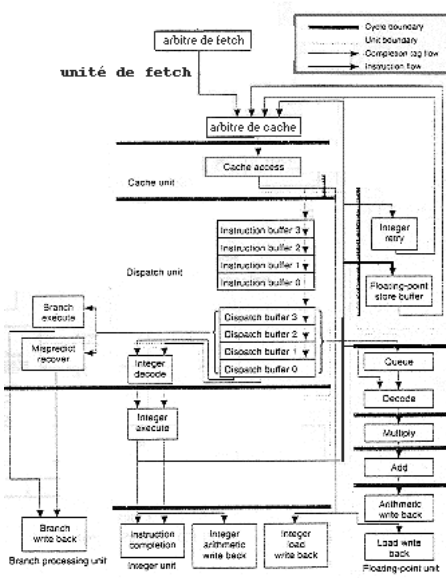
Les adresses de chargement ou rangement de données en mémoire, au préalable dans le cache, sont gérées par la FXU. Le cache possède deux registres tampons d'adresse («integer retry» et «floating point store buffer») qui évitent de par trop bloquer les unités fonctionnelles (FXU ou FPU). Les adresses de branchement sont retournées par la BU directement à l'unité de chargement.

Les autres unités intégrées sur la matrice du PPC 601 concernent essentiellement la gestion de la mémoire et la gestion de l'interface avec le bus processeur. Elles ne font pas partie du noyau fonctionnel du processeur et seront traitées ultérieurement.

Le lecteur familier de présentations de processeurs remarque ici plus encore qu'ailleurs, que l'unité arithmétique et logique, cœur de la machine de Neumann disparaît quasi totalement en tant que telle. Les opérateurs réalisant un jeu complet des opérations d'une algèbre de Boole sont devenus des unités terminales de tuyaux de cheminement plus ou moins spécialisés. Alors que la partie centrale du processeur était cette UAL entourée de registres et de bus internes, cette présentation a complètement disparu.



ARCHITECTURE GENERALE DU POWERPC601



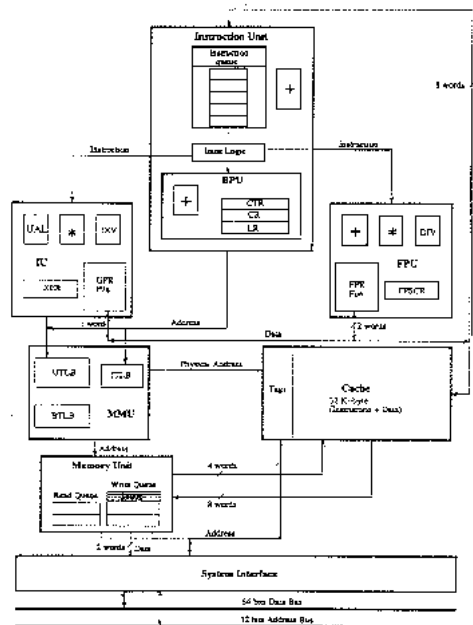
Architecture du pipeline du PPC601

L'unité de fetch (FU) construit les adresses des instructions à charger et les envoie au cache unifié. Le cache, notamment par l'étage d'arbitrage de l'UGM, résout les conflits entre les diverses unités qui accèdent au cache : la FU, les unités fonctionnelles FXU et FPU, mais aussi l'interface avec le bus. Lorsque le chargement d'instructions est autorisé, le registre d'accès au cache est initialisé avec une adresse.

Annexe 3, Architecture RISC

Cette adresse est alors utilisée pour rapatrier 32 octets d'instructions (8 instructions de 32 bits) vers la DU. Celle-ci, en fonction des adresses produites par la FU et de la place disponible dans sa file de stockage à 8 emplacements, reçoit jusqu'à 8 instructions dans le même cycle d'horloge.

Seuls les emplacements du bas de la file sont utilisés pour répartir les instructions entre les unités fonctionnelles. Les instructions peuvent être distribuées dans le désordre, sauf celles destinées à la FXU qui ne peuvent être extraites que depuis le registre du bas de la file. Ceci garantit la distribution des instructions entières dans l'ordre du programme vers une unité qui joue le rôle de chef d'orchestre du noyau du PPC601.



3.2.2 Les unités fonctionnelles du PPC 601

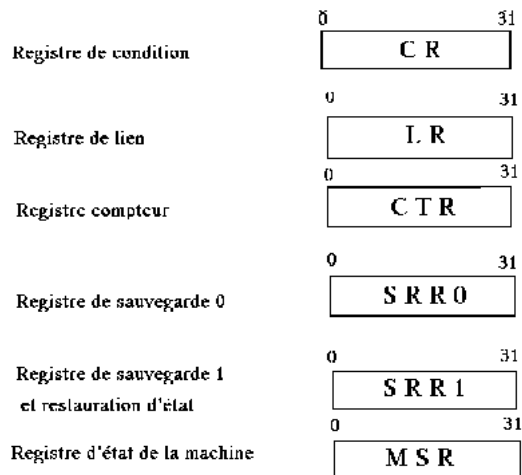
3.2.2.1 Unité de branchement BU pour Branch Unit

En vue logique, la BU groupe les fonctions de chargement des instructions, de leur distribution, et de détermination des adresses de branchement. Dans la réalisation du PPC 601, ces notions sont séparées. La BU du PPC 601 correspond à la seule fonction de choix d'adresse. L'unité de branchement gère les branchements tant conditionnels qu'inconditionnels à l'aide de différents modes d'adressage.

Les registres de la BU

Pour faire ce travail la BU dispose de plusieurs registres de 32 bits :

Annexe 3, Architecture RISC

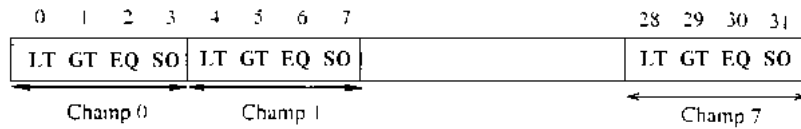


Registres de l'unité de branchement

CR pour Condition Register, registre de condition, est utilisé par les branchements conditionnels. Il est divisé en 8 champs de 4 bits (CRO à CR7) qui portent les conditions d'infériorité, de supériorité, d'égalité et de débordement cumulé, ce dernier est la copie du bit SO du registre d'exception XER de la FXU. Les bits de ces champs sont gérés par des instructions entières ou flottantes.

LR «link register» ou registre de lien peut contenir :

- l'adresse cible de certaines instructions de branchement conditionnel;
- l'adresse de retour d'un branchement à un sous-programme.



(a) Champs du registre de condition.

Bits du CR	Instructions arithmétiques	Instructions de comparaison
LT = 1	résultat < 0	(RA) < (RB), ou (RA) < opérande immédiat
GT = 1	résultat > 0	(RA) > (RB), ou (RA) > opérande immédiat
EQ = 1	résultat = 0	(RA) = (RB), ou (RA) = opérande immédiat
SO	Débordement cumulé	copie de XER(SO) (voir section 7.2)

Registre de condition

CTR pour Count Register, registre compteur a lui aussi deux usages :

- contenir une adresse cible pour un branchement conditionnel;
- gérer le comptage des boucles.

Notons que des registres cachés nommés shadow registers, dédoublent LR et CTR. C'est la technique des registres de renommage qui :

- évite une dépendance de donnée entre une instruction de la FXU qui précéderait une instruction de la BU, celle-ci accédant au même registre;
- évite les incohérences dans l'exécution du programme, dues à la distribution désordonnée des instructions entre les trois unités fonctionnelles;
- aide à sauvegarder et restaurer le contexte pour traiter une exception.

Annexe 3, Architecture RISC

MSR pour «machine state register», registre d'état de la machine, est le registre dont les bits définissant l'état du processeur :

- état de fonctionnement privilégié pour l'exécution d'instructions système;
- interruptions activées ou désactivées;
- suivi en trace des instructions;
- fonctionnement en mode «little endian» plutôt que «big endian».

SRR0 et SRR1 sont des registres de sauvegarde et de restauration (save and restore register). Ils gèrent les changements de contexte pour les interruptions.

Le pipeline de la BU

Le pipeline de branchement est très ramassé. Il est constitué de deux étages, sachant que le dernier est prévu pour être utilisé le plus rarement possible :

«decode register;
execute branch;
predict write back.»

Le premier étage du pipeline est exécuté dans le même cycle d'horloge que la distribution des instructions. Dans le meilleur des cas un branchement en zéro cycle est réalisable.

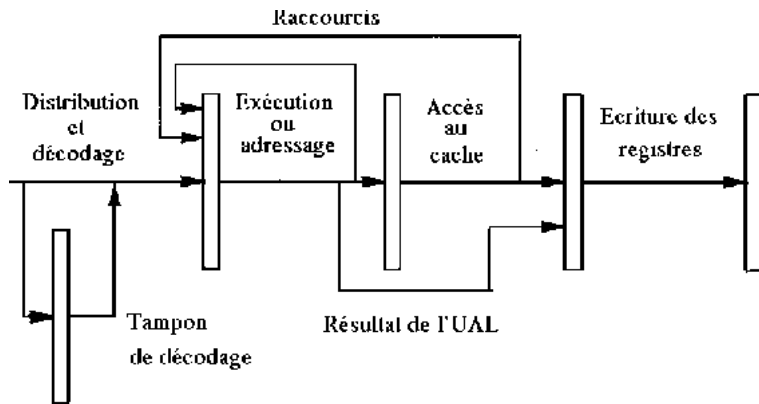
Mécanismes d'optimisation

Le mécanisme de folding, c'est-à-dire l'extraction et la distribution non ordonnée des instructions de branchement, accélère le processus de branchement. Il augmente et régularise le débit d'alimentation des pipelines.

La prédiction de branchement est un autre mécanisme d'accroissement du débit. Il anticipe un branchement conditionnel non résolu en fonction d'une estimation de la probabilité de réalisation d'une condition. Pour ce faire, le PPC 601 utilise un algorithme de prédiction de branchement statique, c'est à dire défini une fois pour toutes à la compilation. Les instructions de la branche prédite sont alors préchargées, voire pré-distribuées. Toutefois, ces instructions ne sont achevées que lorsque la prédiction a été reconnue correcte.

La BU contient aussi un additionneur pour calculer des adresses indépendamment de la FXU ou de la FU et donc améliorer le travail en parallèle de ces unités.

3.2.2.2. Unité entière ou Fixed Point Unit



Pipe line de la FXU

Selon les auteurs, le pipeline entier est vu comme ayant trois ou quatre étages selon que l'on intègre ou non l'accès au cache, en fait à l'UGM qui contient la logique d'accès à ce cache, pour les opérations de rangement ou de chargement.

Le principe de fonctionnement en trois étages est retenu ici. Le premier étage décode l'instruction. Le deuxième exécute le traitement demandé ou le délègue pour partie, soit au cache, soit au séquenceur. Le dernier étage écrit les résultats dans les registres. Ce dernier étage gère aussi la terminaison d'une instruction (completion). C'est là qu'est garanti l'achèvement des instructions dans l'ordre prescrit par le programme.

Les registres

Ils ont tous 32 bits.

Le registre R0 est toujours vu comme valant 0 dans les instructions de chargement ou de rangement. Sinon, c'est un registre de travail de l'unité entière comme les 31 autres.

Annexe 3, Architecture RISC

XER registre d'exception gère certaines exceptions : dépassement de capacité, bit de retenue et gestion des longueurs de chaînes pour les chargements ou rangements multiples.

Le registre MQ, multiplicateur et quotient, qui n'appartient pas à l'architecture PowerPC, existe dans le PPC 601 pour la compatibilité avec l'architecture POWER.

DAR est le registre d'adresse des données.

DSISR est le registre d'état d'interruption.

Ces deux registres servent au système d'exploitation pour gérer les interruptions.

Fonctions de cette unité.

Chargement et rangement :

calcul d'adresses mémoire tant pour les entiers que pour les flottants;

chargement des entiers et des flottants;

rangement des entiers;

opérations arithmétiques et logiques :

addition, soustraction, multiplication, division;

opérations logiques;

opérations de rotation et décalage;

gestion du système :

transfert des registres de travail vers les registres systèmes et vice et versa.

Autres opérations:

opérations particulières de lecture/écriture de registres;

comparaison mettant à jour le registre CR de la BU;

lecture/écriture des registres CTR et LR de la BU;

gestion des dépendances des données et dans le contrôle de flot.

Une des fonctions les plus importantes de la FXU est de gérer les dépendances des instructions afin de garantir la sémantique des programmes. Ces dépendances sont de deux natures :

- dépendances de contrôle de flot;

- dépendances de données.

L'ordre séquentiel du programme est garanti malgré la distribution désordonnée des instructions vers les unités fonctionnelles. Cependant, les instructions sont distribuées et s'exécutent de façon ordonnée à l'intérieur de leur unité respective. Si l'on ajoute à cela les profondeurs différentes des trois pipelines, l'ordre de terminaison des instructions peut être très différent de l'ordre initial. Il est indispensable de garantir la sémantique de l'ordre séquentiel, appelée ici "respect des dépendances de contrôle de flot" entre les unités.

Les dépendances de données nécessitent la cohérence d'accès aux ressources que constituent les registres et la mémoire. Une partie de ces dépendances d'accès provient de l'architecture pipeline et sont donc du type intra-unité . Les autres proviennent du fait que les unités fonctionnelles ne sont pas totalement indépendantes entre elles. Ces dépendances inter-unités sont dues à certains de leurs registres accessibles en lecture ou écriture par les autres unités :

- l'unité entière intervient dans le chargement et le rangement des registres de l'unité flottante;
- l'unité entière accède en lecture/écriture aux registres LR et CTR de la BU qui a aussi les mêmes droits sur ceux-ci;
- les unités entière (FX(T) et flottante (FPU) accèdent en écriture au registre CR de la BU qui est lui-même accessible en lecture par la BU et la FXU.

Les mécanismes de maintien de cohérence et d'optimisation se décomposent en :

.le mécanisme de gestion des dépendances

Les dépendances de données internes à une unité sont gérées par celle-ci. L'unité entière gère en plus les dépendances entre unités, qu'elles proviennent du contrôle de flot ou du conflit d'accès aux ressources. Un système d'étiquetage des instructions entières [J2] fournit cette gestion centralisée des dépendances inter-unités.

Le principe de fonctionnement de ce système d'étiquetage est que chaque instruction flottante ou de branchement susceptible de modifier l'état du processeur produit une étiquette propre au type d'instruction. Par "modifier l'état du processeur", on entend la mise à jour de registres ou une exécution spéculative de branchement.

Chaque étiquette progresse le long du pipeline entier en étant attachée à une instruction entière qui peut recevoir une étiquette de chaque type. L'étiquette est produite lors de la distribution des instructions. Elle est attachée à l'instruction entière libre qui précède l'instruction génératrice. Au besoin, une instruction vide de non opération sera produite et mise dans le pipeline. Une étiquette peut d'ailleurs être détachée d'une instruction entière normale en cours de transfert dans le pipeline et être rattachée à une instruction vide insérée à l'étage que l'instruction support vient de quitter. Ceci bloque les instructions ultérieures, en attendant, par exemple, la résolution d'un branchement, sans pénaliser la terminaison de l'instruction qui la portait.

Les étiquettes sont typées, sans être numérotées. En effet, les instructions s'exécutent dans l'ordre du programme dans leurs unités respectives. Il suffit alors de mettre en correspondance les étiquettes avec les instructions flottantes ou de branchement par comptage des étiquettes en remontant le long du pipeline entier depuis l'étage de terminaison.

Lorsqu'une étiquette arrive à l'étage de terminaison, les mises à jour de registres ou de mémoire correspondantes sont insérées. Ainsi, l'effet de l'instruction sur l'état du processus et de la mémoire est le même que si les instructions s'étaient déroulées dans l'ordre du programme.

Notons qu'il y a ici une optimisation de l'ordre du programme. Les instructions qui ne sont pas susceptibles de modifier l'état de la machine peuvent être exécutées et terminées dans un ordre différent de l'ordre séquentiel.

Annexe 3, Architecture RISC

Ce mécanisme d'étiquetage a été étendu aux mises à jour des 8 champs du registre de condition CR de la BU. L'étiquette est alors à 8 bits, un par champ. Cette étiquette est associée à l'instruction qui fait une mise à jour si c'est une instruction entière. Si l'instruction est une instruction flottante, l'étiquette à 8 bits est associée à une instruction de non opération étiquetée uniquement flottante; cette instruction correspond évidemment à l'instruction flottante de mise à jour du champ du CR.

À la différence de l'étiquetage normal, appelé encore étiquetage de terminaison, les bits de l'étiquette à 8 bits tiennent lieu de sémaphores pour signaler aux autres instructions d'attendre la mise à jour de champs du CR.

La BU ne lira le champ du CR que lorsque toutes les instructions entières et flottantes qui la précèdent dans l'ordre du programme et qui mettent à jour le CR seront terminées. Ajoutons enfin que les exceptions tant synchrones qu'asynchrones sont gérées de façon précise grâce à ce mécanisme d'étiquette. Ce dernier classe les instructions comme étant antérieures ou postérieures à l'exception et les fait terminer ou abandonner en conséquence.

.l'optimisation du pipeline

II convient de signaler :

- le raccourci entre l'étage d'exécution et l'étage de décodage pour l'exécution sans temps mort d'une instruction dépendante du résultat de la précédente;
- le raccourci du cache vers l'étage de décodage qui autorise la lecture d'une donnée sans accès préalable au registre;
- la présence d'un tampon à accès optionnel dans l'étage de décodage pour ôter une instruction entière de la file même si une instruction est déjà en attente dans l'étage de décodage. La distribution des instructions en est accélérée.

.le séquenceur

Cette unité intégrée au processeur peut être vue, pour partie, comme une unité d'assistance à l'étage d'exécution de l'unité entière par le fait que cette dernière lui délègue l'accès en lecture ou écriture à certains registres de contrôle du système. Mais cette unité assure aussi des tâches de bas niveau essentielles au fonctionnement du processeur :

- initialisation du processeur au démarrage;
- gestion de la base de temps;
- gestion des interruptions;
- gestion de la TLB et de l'UGM (cf. gestion de la mémoire).

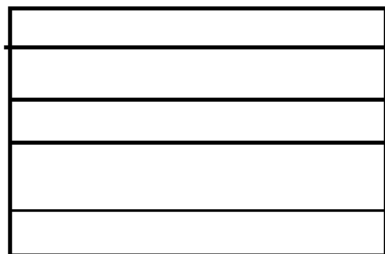
3. Unité de calcul en virgule flottante ou Floating Point Unit

Cette unité fait les calculs en virgule flottante. Tout nombre flottant à charger depuis la mémoire ou à stocker en mémoire l'est par l'intermédiaire de la FXU qui s'occupe totalement du chargement et de la génération d'adresses pour le stockage.

La FPU exécute les opérations de multiplication, division, addition, soustraction en simple et double précisions sur les nombres Flottants. Elle offre la possibilité d'une opération de multiplication-addition combinée à 4 opérandes fréquemment employée dans le calcul matriciel et l'imagerie numérique. Elle est optimisée ici pour le calcul en simple précision ($d = a*b + c$). Le pipeline flottant est spécialement conçu pour cette opération.

Certaines opérations flottantes mettent aussi à jour des registres de la BU, en particulier le CR.

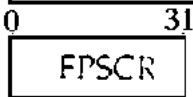
Les registres



Registre de nombre flottant 0

Registre de nombre flottant 1

Registre de nombre flottant 31

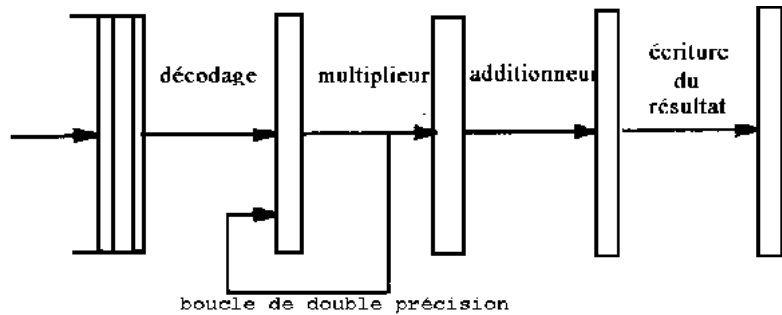


Registre d'état et de commande

Les FPR pour Floating Point Registers, au nombre de 32 et d'une taille de 64 bits, constituent des registres généraux de travail. Le registre FPSCR sert avant tout à signaler les exceptions qui ont pu se produire lors d'un calcul. Une autre de ses fonctions est de gérer la mise en oeuvre facultative de la norme IEEE-754 de l'arithmétique flottante. L'application de cette norme, qui n'a pas été conçue pour les architectures pipeline, diminue les performances de calcul.

Les exceptions en flottant sont précises, elles signalent l'instruction fautive en positionnant 2 bits du registre MSR de la BU (FEO et FE1). Mais imposer une gestion précise des exceptions flottantes diminue les performances de calcul.

Le pipeline de l'unité flottante



Le pipeline est constitué de quatre étages (figure 9). On voit l'architecture est adaptée à l'opération combinée de multiplication-addition. Les instructions simples de multiplication-addition passent au travers d'un étage d'exécution sans utilité réelle (pas de raccourci mis en &oeil;uvre).

Un raccourci créé pour économiser du matériel fait exécuter la multiplication en double précision au prix d'un cycle de latence supplémentaire. Les opérations de division nécessitent ce qui est inévitable, plusieurs passages dans le pipeline, 17 cycles d'horloge en simple précision et 31 cycles en double précision.

Les mécanismes d'optimisation ne contiennent pas de raccourcis de lecture du banc de registres. Une instruction dépendant d'une autre instruction flottante doit attendre que l'écriture des registres soit terminée avant d'être décodée. On peut donc comprendre l'utilité de la file d'attente en entrée du pipeline flottant qui permet au moins d'accélérer la distribution des instructions.

La gestion des adresses de chargement et de rangement de données en mémoire n'est faite que par la FXU. Pour le rangement, la FPU gère toutefois l'écriture dans un registre FPR quand le chargement est, lui, entièrement pris charge par la FXU. Cette gestion du chargement peut permettre d'exécuter dans le même cycle d'horloge, l'écriture d'une donnée du cache dans un registre FPR et le décodage d'une instruction de la FPU dépendante.

La FPU du PPC 601 est moins efficace que celle de l'IBM POWER dont elle découle. Cela est dû à un manque de place sur la matrice du processeur où le cache primaire est, lui, richement doté. Dans cette unité on peut noter l'absence du renommage de registres et celles de nombreux raccourcis (bypass), en particulier ceux de lecture du banc de registres (FPR's). Comme le PPC 601 était prévu pour attaquer le marché du PC où le calcul flottant est un besoin moins important, on peut comprendre la raison du compromis fait sur la FPU.

3. Jeu d'instructions

Le jeu d'instructions du PPC 601 respecte la philosophie RISC en rendant les ressources matérielles visibles pour le compilateur. L'architecture PowerPC a pourtant un jeu d'instructions plus étendu que celui des autres architectures RISC tout en étant très nettement moins important que celui d'une architecture CISC

1. le Format d'instructions

Principaux formats d'instructions du PowerPC 601

Code d'opération	registre destination	registre source 1	registre source 2	extension du code d'op.
------------------	----------------------	-------------------	-------------------	-------------------------

instruction registre à registre

Code d'opération	registre destination	registre source 1	constante
------------------	----------------------	-------------------	-----------

instruction registre - immédiat

Code d'opération	extension du code d'op.	code de condition	constante	extension du code d'op.
------------------	-------------------------	-------------------	-----------	-------------------------

instruction de branchement conditionnel

Code d'opération	constante	extension du code d'op.
------------------	-----------	-------------------------

instruction de branchement inconditionnel

Code d'opération	registre destination	registre source 1	registre source 2	registre source 3	extension du code d'op.
------------------	----------------------	-------------------	-------------------	-------------------	-------------------------

instruction d'addition-multiplication flottante

Le format des instructions est conditionné par les différentes méthodes d'adressage disponibles. Celles-ci sont :

- l'adressage absolu : direct, l'adresse constitue un champ de l'instruction;
- l'adressage indirect : via le registre de base, ce n'est qu'un cas particulier de l'adressage à base;
- l'adressage à base : <registre de base> + immédiat;
- l'adressage indexé : <registre de base> + <registre d'index>.

Une pré-incrémentation est réalisée dans des instructions qui mettent à jour au préalable le registre de base par un calcul d'adresse. Elle est utile pour gérer les boucles de balayage des éléments d'un tableau, l'accès à la mémoire est fait en un seul cycle. Cette technique est rendue possible par l'existence d'un deuxième port d'écriture sur le banc de registre entiers.

Sans conditionner directement le format d'instructions, les types de données reconnus par le processeur interviennent dans sa définition. Ces types de données sont : l'octet, le demi-mot, le mot (32 bits), le double-mot. Ce dernier est utilisé pour le calcul flottant en double précision.

Les instructions de l'architecture PPC ont toutes 32 bits. Elles sont alignées en mémoire. Les instructions ont normalement trois opérandes. Deux champs de l'instruction contiennent les adresses des registres sources et le troisième celle du registre destination. Les instructions combinées de multiplication-addition ont entraîné la création d'instructions à 4 opérandes. Aucun des opérandes n'est en mémoire principale. Toutes les opérations, exceptées celles de chargement et de rangement, sont de registre à registre, conformément aux principes RISC.

On se reportera aux documents spécialisés pour la liste des instructions.

3.4 L'interface du noyau du processeur avec les unités matérielles intégrées ou externes

L'interface gère les accès à la mémoire et les entrées et sorties. Ces dernières sont vues par l'adressage comme une extension de la mémoire segmentée. L'architecture définit en premier lieu un modèle d'espace de mémoire et les algorithmes afférents de gestion d'adresses.

L'architecture PowerPC définit aussi le modèle de mémoire employé, mémoire partagée à ordonnancement lâche conçue pour fonctionner en environnement multiprocesseur. Ceci impose d'avoir des instructions supplémentaires pour gérer la cohérence mémoire d'une part et la cohérence cache/mémoire d'autre part. Ces instructions, utilisées le plus souvent par les systèmes d'exploitation, seront abordées dans les paragraphes correspondants. Cependant, l'architecture PowerPC n'impose pas d'obligation pour l'implantation des caches et de l'interface avec le bus processeur. Ces éléments seront néanmoins présentés car il font partie intégrante de la réalisation du PowerPC 601.

1. Le modèle d'espace de mémoire

Annexe 3, Architecture RISC

L'espace de mémoire défini par l'Architecture PowerPC sera présenté selon ses différents espaces d'adresses hiérarchisés. Les algorithmes de passage de l'un à l'autre seront aussi abordés. Bien que l'architecture PPC n'impose guère de consignes d'implantation en ce qui concerne la gestion et l'optimisation d'accès à la mémoire, les unités concernées du PPC601 : l'UGM et les caches, interne et externe, seront décrits.

3.4.1.1. Espace de mémoire et adressage

I. Espace de mémoire

L'espace de mémoire est constitué de 3 niveaux hiérarchisés.

- *L'espace logique* vu par un programme.

Un processus peut gérer au maximum un espace d'adressage linéaire de 2^{32} octets (4 Go). Cet espace est en fait constitué de 16 segments ou de 4 blocs de mémoire de l'espace virtuel. On parle d'adresse effective plutôt que d'adresse logique dans l'Architecture PowerPC.

- *L'espace virtuel* vu par le système d'exploitation.

C'est un espace segmenté contrairement à celui du processeur Alpha qui est linéaire. Il est constitué de deux types de segments, les segments ordinaires et les blocs de mémoire. L'espace virtuel est de 2^{52} octets soit 4 Petaoctets (4×10^{15}).

Les segments sont constitués de pages de mémoire non nécessairement contiguës. Leur taille est fixe à 256 Mo (2^{28} octets). II ne peut y avoir plus de 2^{24} segments dans l'espace virtuel.

Le bit 0 (bit de poids fort nommé T) de l'adresse de segment (sur 32 bits avec 24 bits d'adresse et 8 bits d'attributs) précise si le segment sert à accéder à la mémoire (T = 0) ou au dispositif d'entrées/sorties. Le bit T permet la compatibilité avec l'architecture POWER.

Deux bits définissent le mode d'accès au segment, superviseur ou utilisateur.

Les blocs sont des segments particuliers au nombre de 4 pour les instructions et de 4 pour les données [L2]. Ils sont constitués de pages mémoire contiguës. Leur taille est paramétrable entre 128 Ko (2^{17} octets) et 8 Mo (2^{23} octets). Comme ce sont des zones de mémoire contiguës, leur système d'adressage peut être simplifié. ils offrent donc un accès rapide en même temps que de larges espaces de travail en mémoire. Si un bloc ou un segment ont la même adresse effective, le bloc sera utilisé. Blocs et segments peuvent se recouvrir.

Deux bits de l'adresse de bloc définissent les accès en mode superviseur ou utilisateur, 6 autres bits d'adresses (W, I, M, G et les deux bits PP) définissent les contrôles d'accès à la mémoire.

- *L'espace physique* constitué par la mémoire principale de l'ordinateur.

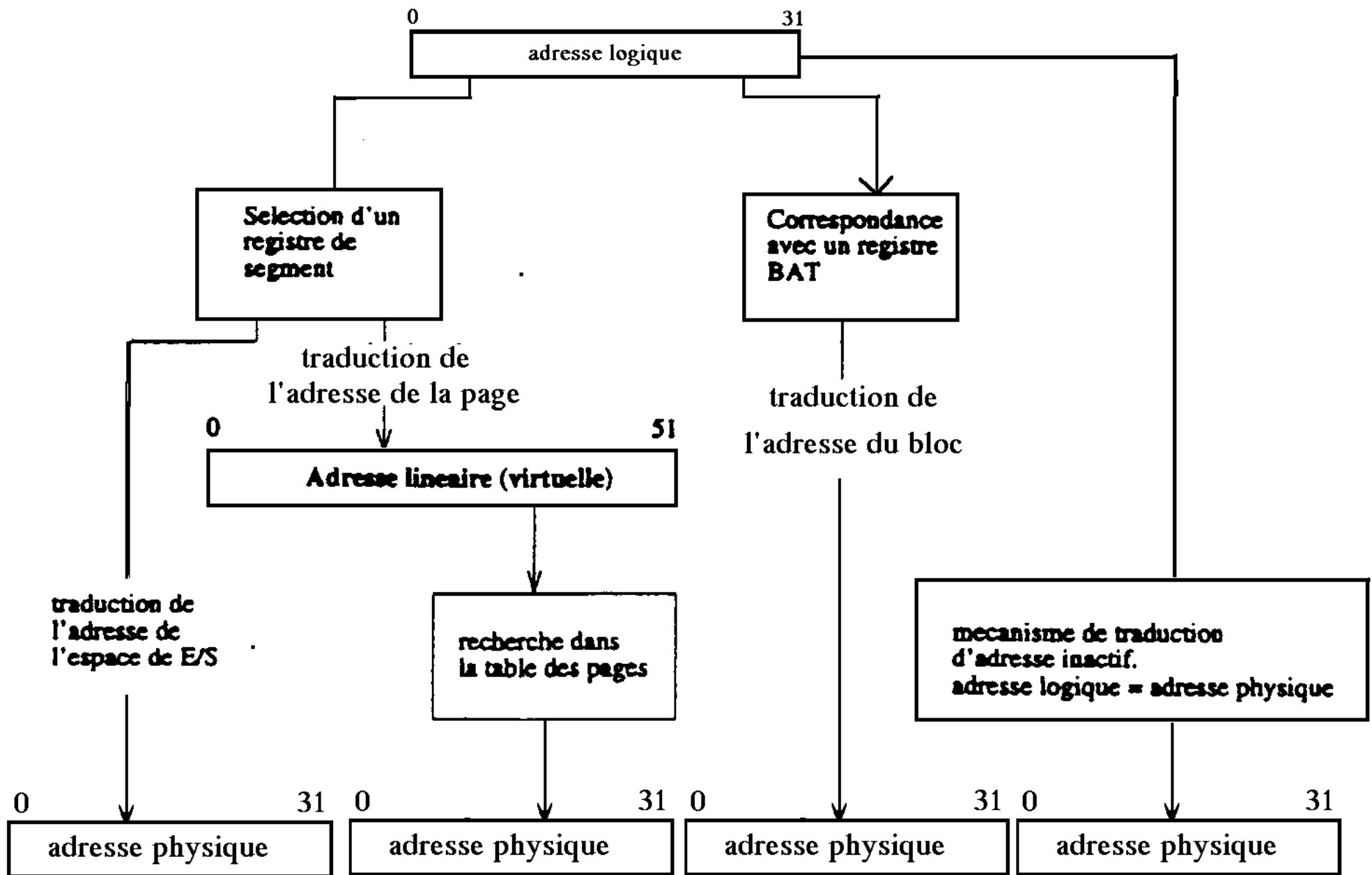
Il est vu comme une table linéaire d'octets d'une taille maximale de 2^{32} octets (4 Go). Il est géré par pagination. Les pages de l'espace physique ont une taille fixe de 4 Ko.

Deux bits de l'adresse de page définissent les accès en mode superviseur ou utilisateur, 6 autres bits d'adresses (W, I, M, G et les deux bits PP) définissent les contrôles d'accès aux pages de mémoire. On parle d'adresse réelle pour l'espace physique.

3.4.1.1.2. Les différents modes d'adressage

1) *le mode d'adressage réel*

Ce premier mode (figure 11) consiste à suspendre tout mécanisme de traduction de l'adresse effective (segmentation et pagination) et à utiliser directement cette dernière pour accéder à la mémoire . Dans ce cas, les accès à la mémoire ne donnent lieu à aucun contrôle, car les protections de la mémoire sont intégrées aux mécanismes désactivés"



mécanismes de traduction d'adresse

2) l'adressage de bloc

Traduction d'adresse de bloc (BAT)

Ce deuxième mécanisme correspond à l'adressage d'un bloc mémoire. Chaque bloc est décrit par deux registres. Le premier correspond à l'adresse effective de début de bloc: le deuxième à l'adresse réelle de début de bloc. Conceptuellement, l'adresse réelle d'un accès dans un bloc est obtenu en soustrayant à l'adresse effective l'adresse effective de début de bloc et en ajoutant l'adresse réelle de début de bloc au déplacement ainsi obtenu.

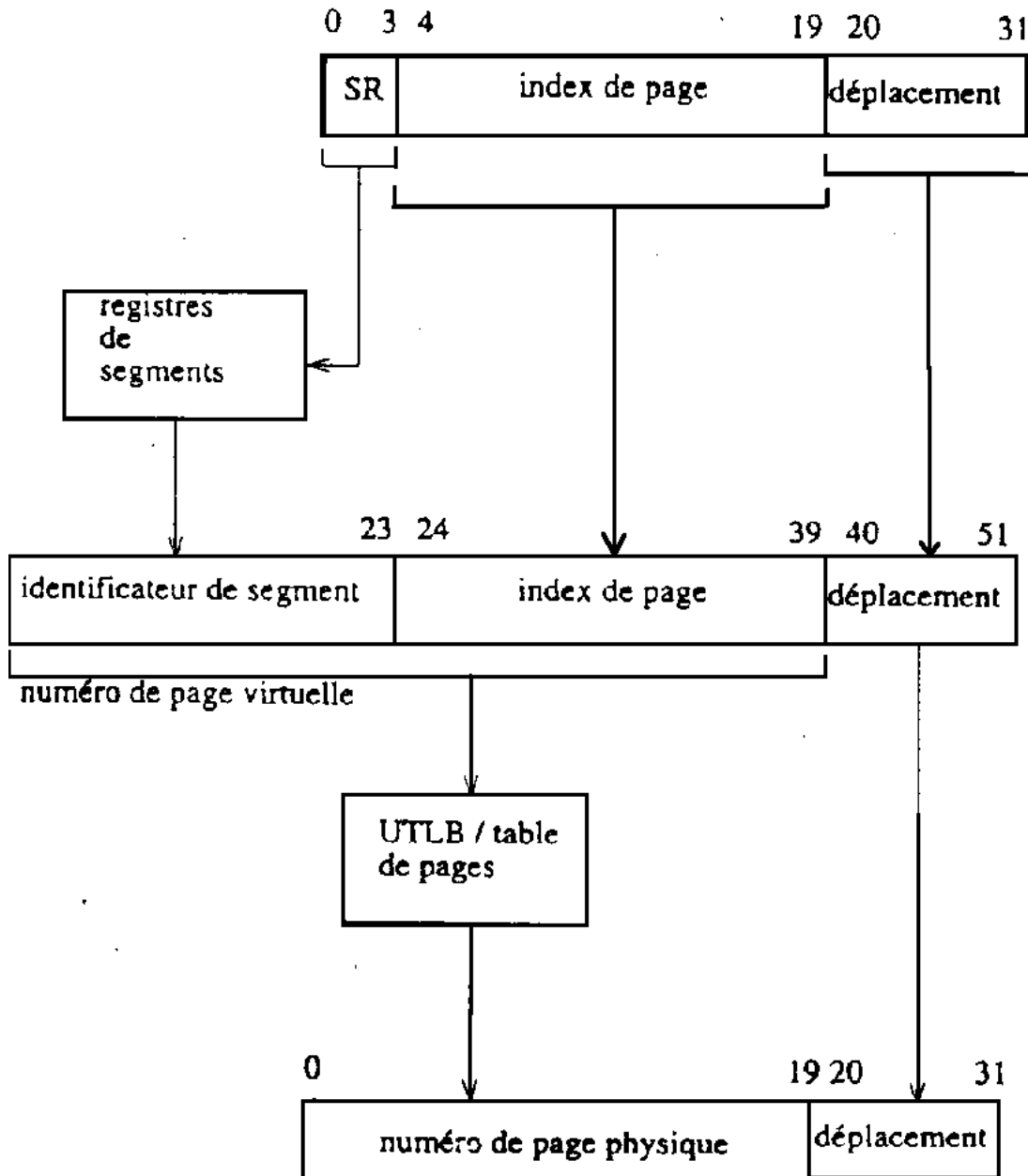
Annexe 3, Architecture RISC

Les deux derniers mécanismes commencent par traduire l'adresse de segment. Les 4 bits de poids fort (bits 0 à 3) de l'adresse effective adressent les 16 registres de segment.

Les segments sont à classer entre les segments d'adressage direct (T=1) et les segments ordinaires.

segment d'adressage direct (T=1):

"un segment d'adressage direct est la projection d'adresses effectives sur un espace d'adressage externe, de type Bus I/O. Les adresses effectives à partir des segments d'adressage direct ne franchissent que la première étape de traduction d'adresse segmentée". [L2 p 421]



Traduction de l'adresse d'une page

Avec un **segment d'adressage ordinaire**, le mécanisme de traduction de l'adresse de segment traduit l'adresse effective en adresse virtuelle laquelle est obtenue en concaténant les 24 bits d'adresse contenus dans l'un des seize registres de segment aux 28 derniers bits (bits 0 à 31) de l'adresse effective. L'adresse virtuelle ainsi obtenue peut alors être traduite en adresse réelle. 6×10^{23}

La HTAB, table des pages en hash, définit les pages de la mémoire virtuelle en correspondance avec celles de la mémoire réelle. Elle est constituée de descripteurs de page (Page Table Entry). Ceux-ci contiennent un VPN (virtual page number), sous forme de l'identificateur de segment et d'un index de page abrégé, et le numéro de page réelle correspondant. Ces PTE sont regroupés par huit, les groupes ainsi formés étant identifiés par une clef de hachage primaire ou secondaire, appelée PTEG (Page Table Entry Group).

À l'aide de l'adresse de page virtuelle (VPN) (bits 0 à 39 de l'adresse virtuelle) et de deux fonctions de hachage donnant accès à deux PTEG, on accède à deux listes de PTE à parcourir jusqu'à obtenir la concordance avec le VPN d'un PTE. La description complète du fonctionnement de l'algorithme d'accès à la table des pages est dans [L2 Vol III par. 4-5-2).

La taille minimale conseillée dans [L2 Vol III] pour la HTAB, est régie par : "Le nombre de PTEG de la table des pages doit être au moins de la moitié du nombre de pages réelles que l'on peut atteindre".

Cette règle est minimale. La taille de la HTAB peut être beaucoup plus grande en cas de risque d'avoir trop de conflit de hash ou plus de seize descripteurs de pages pour la même paire clef primaire/clef secondaire de PTEG. Evidemment, par trop agrandir la table signifie que beaucoup de PTE ne seront pas utilisés ...

2. L'unité de gestion de cette interface : L'UGM

L'UGM, Unité de Gestion de la mémoire (MMU ou Memory Management Unit) interface le noyau fonctionnel du processeur et la mémoire sous ses divers aspects :

- le cache primaire unifié intégré;
- la mémoire externe, cache de deuxième niveau ou mémoire de travail, atteinte au travers de deux unités intégrées au processeur : un tampon de mémoire (File d'attente de la mémoire) et l'Unité d'Interface de Bus (UIB).

Les rôles de l'UGM sont :

- de gérer les concurrences d'accès entre l'unité de chargement d'instructions et la FXU laquelle gère l'accès aux données mémoire, au moyen d'un étage d'arbitrage lequel conduit à stocker l'adresse de donnée ou d'instruction sélectionnée dans le registre d'accès au cache;
- de gérer la concurrence d'accès au cache à la fois par le coeur du processeur et l'interface du bus;
- de traduire les adresses effectives (ou logiques) passées au noyau en adresses virtuelles puis en adresses réelles pour préparer un accès à la mémoire externe, ceci en utilisant les caches de traduction d'adresse (UTLB et TSA) si possible, en parallèle avec cette traduction d'adresse;
- de comparer les étiquettes de l'UTLB (de la TSA en premier lieu s'il s'agit d'instructions) et celles du répertoire de cache avec l'adresse effective afin d'accéder si possible aux données ou instructions présentes dans le cache;
- de passer l'adresse réelle à la File d'attente de la mémoire, dans le cas d'accès infructueux au cache ou de mise à jour de la mémoire externe depuis le cache,.

Il faut noter que l'UGM est couplée à la fois à l'unité de chargement d'instructions et à la FXU au point de faire apparaître l'accès au cache (lorsqu'il n'y a pas d'interblocage) comme un étage supplémentaire entre chargement et distribution d'une part et les étages d'exécution et de terminaison/écriture dans les registres d'autre part.

Cependant, dans les cas favorables, l'unité de chargement d'instructions libère l'UGM de la traduction d'adresse effective grâce à une copie partielle de la UTLB appelée Translation Shadow Array.

Les calculs d'adresses de données (par la FXU) et d'instructions peuvent alors être faits en parallèle. Cependant l'accès au cache ne se fait que par un seul port pour les deux unités. Aussi l'étage d'arbitrage de l'UGM reste-t-il employé pour résoudre cette dépendance et l'une des demandes d'accès reste en attente au mieux pendant un cycle.

L'algorithme d'arbitrage donne priorité aux accès aux données par l'unité entière et aux échanges entre la mémoire et le cache. La localisation de la TSA est sujette à interprétation :

- dans l'unité de chargement;
- dans l'UGM.

La première solution semble plus logique pour les calculs des adresses en parallèle, elle est confirmée par deux sources différentes.

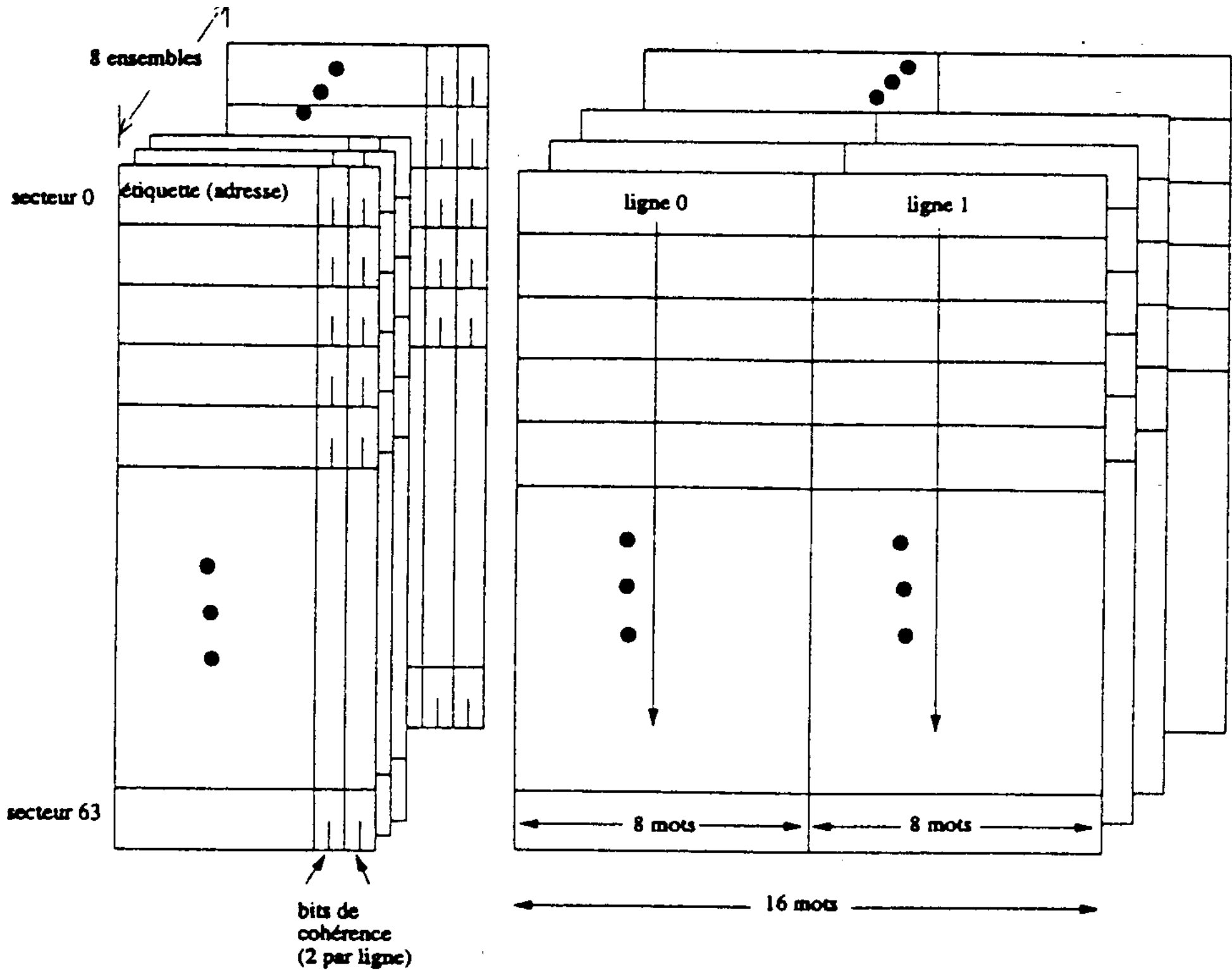
Pour accélérer l'obtention de l'adresse réelle par lecture de la table des pages sise en mémoire principale, un cache de la table des pages est implanté dans l'UGM. Il traduit les adresses des instructions et des données, d'où son nom : UTLB (Unified Translation Lookaside Buffer). Il a une organisation associative en deux ensembles. Il est mis à jour automatiquement par le séquenceur.

3. Le cache primaire unifié intégré

L'organisation retenue pour le cache primaire est une organisation associative par 8 blocs. C'est à dire que le cache est divisé en 8 blocs d'emplacements et que chaque donnée de la mémoire principale ne peut être chargée que dans l'ensemble dont l'indice est fonction de la place de la donnée en mémoire [R1 p60).

Chaque ensemble est constitué de 64 lignes de caches de 64 octets appelées blocs dans la terminologie du PPC. Chaque bloc est divisé en deux secteurs de 32 octets. Le secteur est l'unité de transfert entre le cache et la mémoire principale.

IMAGE POWPC12



Optimisation du cache du PowerPC 601

Le cache est indexé par l'adresse réelle (physique) et les étiquettes de blocs sont associées de même à l'adresse réelle. Le cache possède :

- un port de lecture de données de 32 octets utilisé concurremment par la FXU, la FU et l'interface de bus pour la réécriture en mémoire;
- un port d'écriture de 16 octets utilisé lui aussi concurremment par la FXU, la FPU et l'interface de bus;
- un port d'adresse utilisé seulement par l'interface de bus pour les opération de lecture/écriture en mémoire.

L'UIB peut alors, profitant de la correspondance directe entre étiquette de cache et adresse réelle, espionner le bus en parallèle avec tout autre accès au cache par le noyau du processeur et ceci sans passer par l'UGM.

Pour comprendre le mécanisme de consultation du cache, il faut prendre en compte quelques données. L'adresse effective est sur 32 bits. Les 6 derniers bits (26 à 31) servent à se positionner dans le bloc de cache. Les 6 bits précédents (20 à 25) servent à se positionner sur le bon bloc du cache. Ces 12 derniers bits de l'adresse effective correspondent comme il a déjà été vu, au déplacement dans la page réelle et ne sont jamais traduits lors d'un accès à la mémoire principale. Il en est de même dans le mécanisme d'adressage du bloc mémoire. Il reste à choisir entre 8 blocs (8 étiquettes) pris chacun dans un ensemble différent mais ayant la même position relative (même adresse relative) dans leur ensemble respectif.

Le choix de l'ensemble où rechercher le bon bloc de cache est fait au départ par le même mécanisme que celui d'accès à la page réelle en mémoire. Il permet ainsi de récupérer l'adresse de page réelle (RPN). Un numéro de page réelle ne pouvant correspondre qu'à un seul ensemble, la comparaison des 8 étiquettes et RPN permet de déterminer l'ensemble à retenir.

On voit ici que le fait d'avoir choisi une taille d'ensemble identique à celle d'une page donne une correspondance aisée entre la mémoire et le cache.

3.4.1.4. Le cache secondaire externe (encore appelé cache de niveau 2 ou L2)

La UIB du PPC601 dispose de trois broches en sortie pour faire savoir au cache L2 quel est l'ensemble du secteur de cache affecté par un transfert. Il est conseillé de prendre pour cache secondaire un cache à correspondance directe incluant le cache primaire L1. La correspondance directe permet de réduire le temps de latence des accès à L2. Correspondance directe signifie qu'à la position d'une donnée (ou un secteur ici) dans la mémoire principale correspond une position unique dans le cache. L'inclusion de L1 dans L2 signifie que toute adresse (et donc donnée) de L1 est aussi gérée par L2. Ceci permet au cache L2 de filtrer ses échanges avec L1 aux seuls transferts de mémoire qui concerne ce dernier. La détermination d'une adresse en mémoire par le processeur permet ainsi au cache L2 d'accéder immédiatement au bon secteur et de délivrer les données en parallèle avec la vérification d'étiquette, quitte à interrompre le transfert vers le processeur si l'accès à L2 est un échec (miss).

Bibliographie

Livres :

[L1] *Shlomo Weiss et James E. Smith, POWER et PowerPC : principes. architecture, réalisation*, International Thomson Publishing France 1995

[L2] *IBM, Architecture PowerPC : spécifications pour une nouvelle famille de processeurs RISC*, International Thomson Publishing, France, 1995

Rapports :

[R1] *Seznec, Vauléon, Étude comparative des architectures des microprocesseurs Intel Pentium et PowerPC601*, INRIA R2320 Août 1994.

Articles de journaux ou de revue :

On consultera utilement **Communications of the ACM** Juin 94, pour six articles.

Diefendorff, History of the PowerPC architecture.

Burgess, Ullah, Van Overen, Ogden, The PowerPC 6U3 microprocessor.

Suessmith, Paap, PowerPC 6U3 microprocessor, power mangement.

Poursepanj, The PowerPC performance modeling methodology.

Shipnes, Phillip, A modular approach to Motorola PowerPC compilers.

Anderson, An overview of Motorola's PowerPC simulator family.

[J1] *BYTE* 11-93 p 209 *NT Roars on the 604*

On consultera également *IEEE MICRO* Octobre 1996, pour plusieurs articles dont :

[J2] *Terence Potter, Mike Vaden (IBM) et Jerry Young, Nasr Ullah (Motorola), Resolution of data and control-flow dependancies in the PowerPC601*

Articles de conférence :

IEEE compcon printemps 93 :

[C1] *Allen, Becker, Multiprocessing aspects of the PowerPC 601.*

Annexe 3, Architecture RISC

[C2] Moore, **The PowerPC 601 microprocessor.**

Paap, **PowerPC : A performance architecture.**

IEEE compcon printemps 94:

Burgess, Alexander, Ho, Litch, Mallick, Ogden, Park, Slaton, **The PowerPC603 microprocessor : a high performance, low power, superscalar RISC microprocessor.**

Steel, **The PowerPC reference platform and machine abstractions.**

Shipnes, **Targeting the Motorola RISC compiler for the PowerPC architecture.**

Holle, **The Windows NT operating system on a PowerPC microprocessor.**

IEEE compcon printemps 95 :

Bannon, Keller, **Internal architecture of Alpha 21164 microprocessor.**

Nicholson, **The RISC system/6000 SMP system.**

Ogden, Kuttana, Loper Mallick, Putrino, **A new PowerPC microprocessor for low power computing systems.**

Levitan, Thomas, Tu, **The PowerPC 620 microprocessor: a high performance superscalar RISC microprocessor.**

Howard C. Thamm, **Developping Windows NT applications for PowerPC microprocessor based systems.**

Clark, Quin, Weaver, **Symmetric multiprocessing for the AIX operating system.**



Conservatoire national des arts et métiers

Architectures des systèmes informatiques

ANNEXE 03

Architecture RISC, exemple du PowerPC

Année 2002-2003

Conservatoire national des arts et métiers
Architectures des systèmes informatiques
 ANNEXE 04
Gestion de la couleur
 Année 2002-2003



L'auteur est GRANDEMENT REDEVABLE à Mme Hélène RAVASIO, candidate reçue à l'oral probatoire dans le CENTRE ASSOCIÉ DU LANGUEDOC-ROUSSILLON (année 1997).

1. Introduction

Le lecteur pourra s'étonner de lire une annexe sur le traitement de la couleur dans un cours relatif à l'architecture des systèmes informatiques. Bien que la colorimétrie et la gestion de la couleur n'en fasse pas partie stricto sensu, les besoins en matériel banal ou spécialisé pour le traitement de la couleur en font un grand consommateur de biens architecturaux. L'augmentation des besoins des utilisateurs demande des modifications architecturales dont on voit les prémisses dans les bus et les mémoires spécialisés.

La couleur est présente en informatique et pas seulement en infographie. Quelle direction financière oserait présenter à son conseil d'administration des graphiques imprimés en noir et blanc ? même pour annoncer des bénéfices, sauf dans une administration; il est vrai que chez elle il n'y a pas de bénéfices.

Lorsqu'on numérise un document en couleurs, qu'on l'affiche à l'écran pour le cadrer ou pour le retoucher, et qu'enfin on l'imprime, on est intrigué ou déçu par les résultats. L'image peut perdre de sa qualité dans la finesse, dans les teintes ou en définition, et le plus souvent elle est modifiée. Ces différences procèdent au moins des causes suivantes.

Image imprimée	Image affichée (vidéo)
Mesure en pouces	Mesure en nombre de pixels
La taille de l'image ne change pas avec sa résolution	La taille de l'image change avec sa résolution
La taille de l'image est modifiée par le facteur d'échelle	La taille de l'image est modifiée par rééchantillonnage
Les pixels de l'image sont séparés en fonction de la résolution	Les pixels de l'image sont localisés un à un sur les pixels de l'écran
On utilise plusieurs points d'encre pour représenter la couleur d'un point	Un pixel d'écran contient un pixel d'image et a la couleur voulue

Le traitement de la couleur par les techniques informatiques est complexe et imparfait.

Dans la suite, on définira d'abord la notion de couleur. On citera quelques bases de la colorimétrie. On abordera ensuite les techniques employées en informatique dans les étapes de manipulation des images en couleurs (fixes ou animées), saisie, mémorisation, affichage, impression, transport.

Il ne s'agit pas d'un cours de traitement des images.

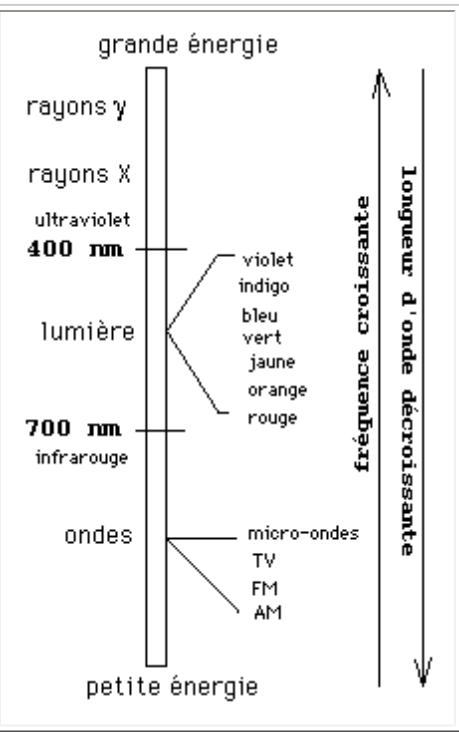
Un mot tout d'abord sur la lumière. C'est une forme de l'électromagnétique ondulatoire qui a un aspect corpusculaire. Les photons se déplacent à environ 300 000 km/s dans le vide. Comme tout phénomène ondulatoire, elle a deux caractères : sa fréquence ou sa longueur d'onde et son intensité. La **fréquence** de la lumière est de l'ordre de six cent mille GHz. Les longueurs d'ondes visibles vont de 400 nm à 700 nm (nanomètres).

Les **rayons gamma** sont produits principalement par des bombardements de protons ou neutrons dans des réactions nucléaires. La fission du noyau crée une excitation de l'atome. L'atome se libère de l'énergie excédentaire en émettant des radiations gamma.

Les **rayons X** sont produits principalement en bombardant une substance par des électrons. On excite ainsi les électrons des premières couches électroniques de l'atome qui produisent de puissantes radiations.

Les rayonnements **ultraviolet, visible et infrarouge** sont émis par la désexcitation des électrons des dernières couches électroniques. Ce type de rayonnement est produit principalement par une stimulation thermique et électrique.

Les **micro-ondes** et les **ondes radio** sont produites par l'excitation électrique des électrons libres d'un conducteur.



2. La couleur et la colorimétrie

La perception de la couleur par l'œil et le cerveau humain est un phénomène complexe qui reste en partie inexpliqué. Néanmoins, certaines théories plus ou moins discutées servent de base à la science qu'on appelle colorimétrie, dont les applications sont multiples dans l'artisanat, l'industrie, la publicité, et bien sûr le secteur informatique.

2.1. Historique des principales découvertes relatives à la couleur

Dans l'antiquité *Platon* reconnaît la couleur comme «impression sensible», il la définit comme «une sorte de flamme qui s'échappe des corps pour s'unir à la vue en produisant la sensation». Pour *Aristote*, elle est une juxtaposition du blanc et du noir.

En 1666, *Newton* réalise une expérience capitale. Il décompose la lumière blanche avec un prisme et fait apparaître sur un écran blanc, un rectangle allongé coloré aux couleurs de l'arc en ciel. On le nomme *spectre solaire* rouge, orange, jaune, vert, bleu, violet, ou encore comme on l'apprend aux enfants : violet, indigo, bleu, vert, jaune, orangé, rouge.

Bien vite, on se rend compte que ce ne sont pas seulement 6 ou 7 couleurs qui entrent en jeu. Notre œil distingue un nombre considérable de couleurs. Cependant il paraît impossible que notre rétine puisse contenir un jeu de récepteurs pour chacune. En 1801, *Thomas Young* énonce la *théorie trichromatique*. Il suppose que la rétine de l'œil humain possède trois sortes d'éléments sensibles respectivement au bleu-indigo, au vert-jaune et au rouge. Il est n'est pas cru. *Helmoltz* fera reconnaître cette théorie cinquante ans plus tard.

Maxwell (1831-1879) fait des expériences, notamment celle de la boîte à couleurs, montrant que *n'importe quelle couleur peut être obtenue par le mélange de trois couleurs de base*. Ceci fonde la *colorimétrie* dont l'objet est de remplacer la vision subjective par un système de mesure objectif et fiable.

Grassman (1809-1877) expose trois lois énoncées plus loin qui sont la base des systèmes colorimétriques.

Artistes et hommes de science ont proposé diverses méthodes de notation systématique pour le repérage des couleurs. *Munsell* (1858-1918) invente le cercle chromatique encore largement utilisé, il est fondé sur des écarts chromatiques visuellement comparables.

Depuis 1931, la *Commission internationale de l'éclairage* (CIE) formée d'experts de différents pays est

l'organisme de référence pour la colorimétrie. Il existe aussi le Comité international de la couleur et le Comité français de la couleur. Ces instances publient des normes.

2.2. Essai de définition de la couleur

La couleur est une composante essentielle de notre perception visuelle. Nous nommons les couleurs avec des termes subjectifs : bleu pâle, verdâtre, bleu vert ou vert bleu etc., ou encore par référence à des couleurs d'objets : orange, marron, turquoise, etc.

La sensation que nous appelons *couleur* d'une surface est le résultat de la réflexion par cette surface d'une partie de lumière blanche et de l'absorption du reste.

Regarder un objet et en décrire la couleur se fait en trois étapes :

- l'étape *physique*, la partie de la lumière réémise par l'objet est sa *couleur physique*. Reçue par l'œil, elle est le *stimulus physique*;
- l'étape *physiologique*, l'œil transforme ce stimulus en signaux nerveux interprétables par le cerveau;
- l'étape *psychologique*, le cerveau interprète ces signaux par référence à sa mémoire correspondant à un ou plusieurs termes descriptifs, bleu, rouge, sombre, clair, pâle, etc.

Pour comprendre les questions liées à la lecture, à la mémorisation et à la restitution des couleurs en informatique, un peu de physiologie de la couleur est utile.

2.3. La physiologie de la couleur

Notre œil a deux parties principales :

- une *optique* qui comprend la *cornée*, le *corps vitré*, la *pupille*, le *cristallin* et les *muscles ciliaires*;
- la *zone de détection*, la *rétine*.

La *rétine* est une mosaïque de cellules photosensibles de 0,3 mm d'épaisseur environ qui distinguent les couleurs. Elle couvre environ les deux tiers de la surface interne de l'œil.

Les cellules photosensibles sont de deux types :

- les *bâtonnets* sensibles aux éclaircissements faibles. Ils ne distinguent pas les couleurs. Ils sont sensibles à l'intensité de la lumière sous les formes noire, grise et blanche;
- les *cônes* fonctionnent en éclaircissement normal ou intense. Ils distinguent la teinte d'un objet observé. Trois types de cônes existent, pourvus de trois pigments visuels différents qui détectent le *rouge*, le *vert* et, en moindre proportion, le *bleu*. La sensibilité de ces cônes est à son maximum que pour une luminosité de l'objet éclairé voisine de la lumière du jour. Il faut donc prendre soin de la lumière qui éclaire un objet pour observer les couleurs de l'objet ou d'une de ses restitutions.

On considère que l'œil d'un homme en bonne santé a les mêmes caractéristiques que celui de la moyenne de la population. Cet observateur normal définit l'ensemble de ses perceptions colorées avec trois couleurs. Il est l'individu trichromate normal. Des variations individuelles apportent des différences de perception des couleurs. Les deux grands cas pathologiques sont les dichromates ou daltoniens qui ne voient pas certaines couleurs et les monochromates qui voient tout en noir et blanc.

La *colorimétrie* a pour objet de mesurer les couleurs, en les rapportant à un *observateur de référence*, personnage fictif censé représenter la moyenne des sujets normaux dans des conditions précises.

2.4 La trichromie

2.4.1. Couleurs fondamentales, primaires, secondaires

Les teinturiers savent qu'on peut reproduire à peu près toutes les teintes en mélangeant trois couleurs

principales : le rouge, le bleu et le jaune. On pensait que cette triade était dans la nature même de la lumière. Young a découvert que le trichromatisme est propre à la rétine. Il est physiologique en ce que chacune de ces trois couleurs excite un des trois types de cônes de l'œil humain.

En fait, suivant les applications, la colorimétrie va utiliser *trois couleurs primaires* au moins. Les couleurs *secondaires* s'obtiennent par mélange de deux couleurs primaires à doses égales.

Exemple :

bleu + rouge -> magenta

rouge + vert -> jaune

vert + bleu -> cyan

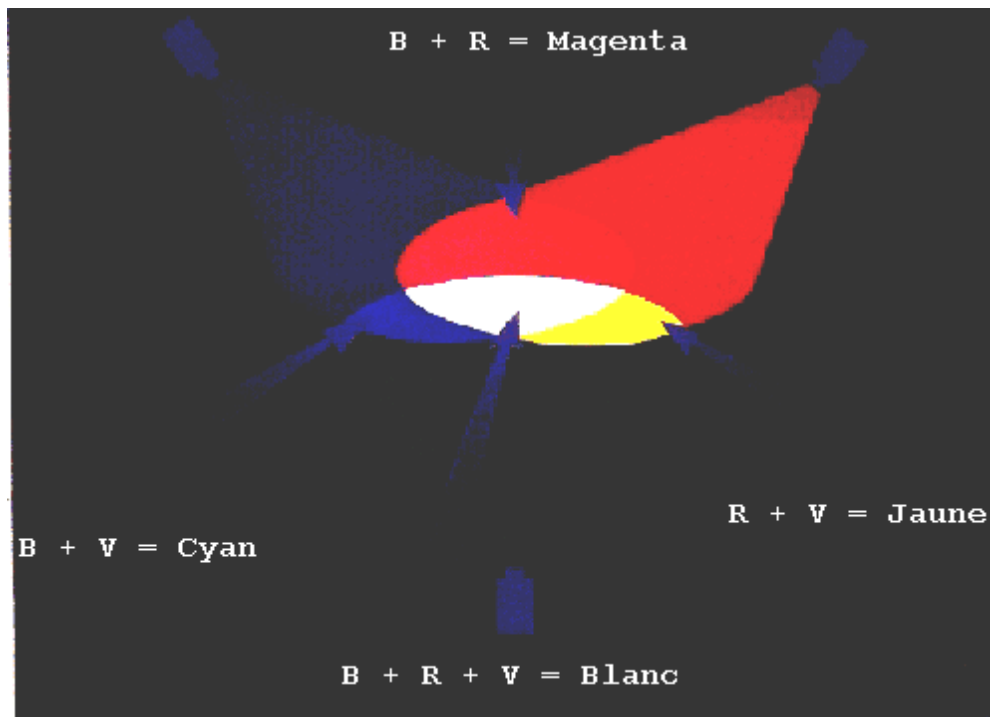
Le blanc est la résultante de toutes les couleurs.

Le noir est l'absence de couleur.

Deux couleurs sont *complémentaires* si elles donnent l'impression d'une lumière blanche en agissant simultanément sur l'œil moyen.

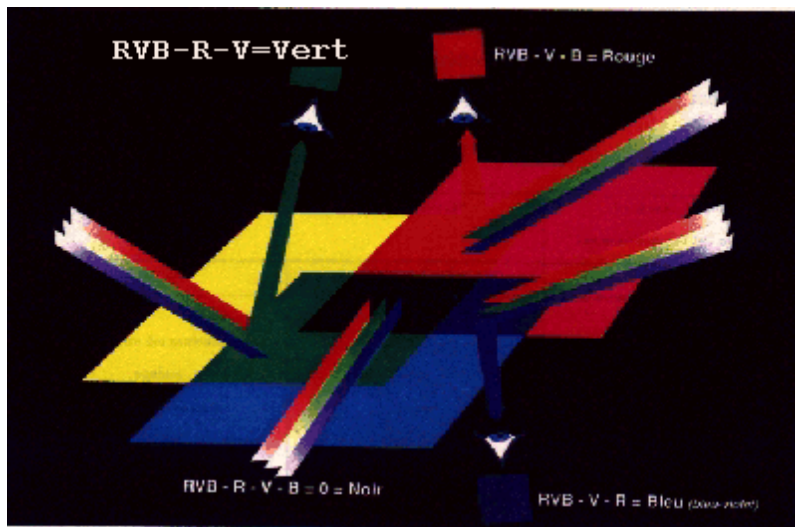
2.4.2. Synthèse additive

Toute couleur visible peut être décomposée en trois couleurs dites primaires. En faisant varier leurs intensités, on produit l'ensemble des couleurs visibles. Si les couleurs désirées sont composées à l'aide de trois faisceaux lumineux fournis par des lampes équipés de filtres colorés, on parle de *synthèse additive*. Dans ce cas, les trois couleurs fondamentales sont le rouge, le vert et le bleu. La lumière blanche est l'addition de ces couleurs avec la même intensité.



2.4.3. Synthèse soustractive

En superposant trois filtres colorés avec une source unique de lumière blanche, on parle de *synthèse soustractive*. Chacun des filtres «retient» une couleur. Les trois couleurs primaires sont ici *cyan, magenta et jaune*.



Chaque surface colorée, éclairée en lumière blanche, absorbe les couleurs de sa couleur complémentaire et diffuse le reste. La surface verte, composée d'une couche jaune superposée à une couche cyan, absorbe les radiations rouges par les pigments cyan et les radiations bleues par les pigments jaunes, seules les radiations vertes sont diffusées et vues par l'œil. Plus il y a superposition de couleurs, moins il y a de diffusion de lumière jusqu'à obtenir du noir.

2.4.4. Caractéristiques physiques d'une couleur

Une couleur est définie par trois grandeurs physiques, bases de la colorimétrie.

2.4.4.1. La teinte ou tonalité chromatique ou espèce

Elle définit la nature de la couleur (bleu, jaune etc.). Un rayon lumineux est une onde électromagnétique, mesurée par sa longueur d'onde. Ce qui est interprété comme une couleur particulière est provoqué par des rayons lumineux de longueurs d'ondes différentes réfléchis par la surface des objets. On mesure la teinte par la longueur d'onde dominante. Les différentes radiations sont classées comme suit :

Radiations	Couleur perçue	Longueur d'onde en nanomètres
ultraviolets	non visibles par l'œil humain	de 1 à 380
visible par l'œil humain	violet	de 380 à 436
Idem	bleu	de 436 à 495
Idem	vert	de 495 à 566
Idem	jaune	de 495 à 566
Idem	orange	de 566 à 589
Idem	rouge	de 627 à 780
infrarouges	non visibles par l'œil humain	de 780 à 1000

2.4.4.2. La pureté ou saturation

Elle indique en quoi la couleur considérée s'approche de la couleur pure correspondante. Cette mesure est égale à 1 pour une lumière monochromatique, couleur *pure* ou *saturée*. Elle diminue à mesure que la couleur est de plus en plus *lavée de blanc*. Le blanc a une pureté nulle. La saturation est un rapport d'énergies, énergie de la longueur d'onde de la tonalité pure divisée par l'énergie totale.

2.4.4.3. L'intensité ou luminosité ou luminance

C'est la quantité totale d'énergie du spectre de la couleur. Plus la couleur est claire, plus son intensité est importante. La luminosité du blanc est 1. La luminosité du noir est 0.

L'espèce et la pureté regroupées sous les termes de *chrominance* ou *chromaticité* de la couleur, définissent la qualité de cette couleur.

L'intensité de la couleur est aussi nommée *quantité* de la couleur.

2.4.5. Lois de Grassman

Grassman a établi les trois lois de la colorimétrie comme une algèbre.

Loi d'additivité : l'intensité d'un mélange de couleurs est la somme des intensités des couleurs composantes.

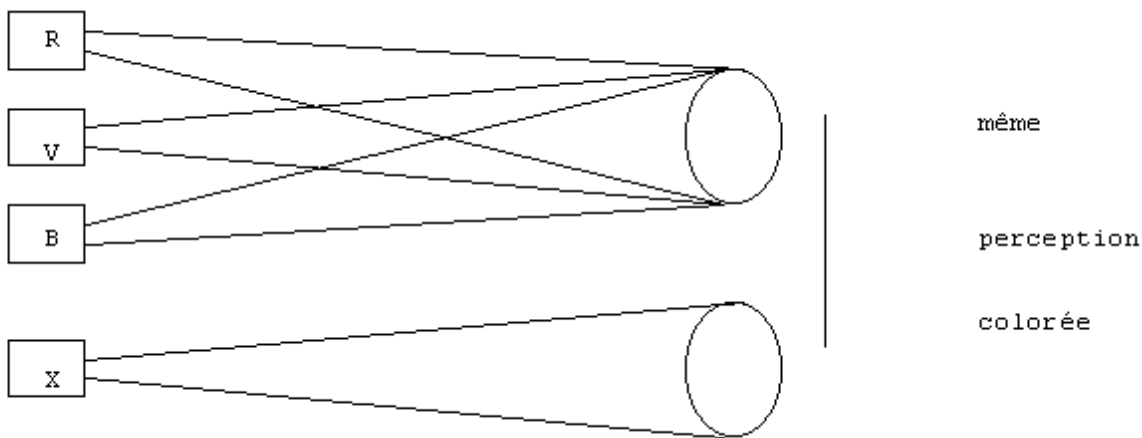
Loi de proportionnalité : si deux plages lumineuses sont perçues comme de même couleur, l'égalité subsiste lorsque leurs intensités sont multipliées par le même nombre.

Loi du mélange des couleurs : deux mélanges lumineux qui, juxtaposés, provoquent la même perception colorée se comportent identiquement dans le processus de mélanges.

2.4.6. Quantification d'une couleur par trichromie

Nous avons vu qu'à partir de trois projecteurs rouge R, vert V et bleu B, on peut constituer une couleur donnée X. La proportion de chaque couleur primaire peut être définie par l'intensité de son projecteur.

intensités



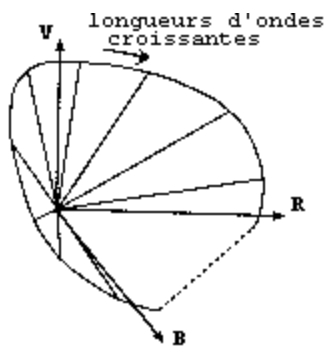
D'après les lois de Grassman, on peut écrire :

$$\mathbf{X} = r(\mathbf{R}) + v(\mathbf{V}) + b(\mathbf{B}).$$

Pendant certaines couleurs très pures sont impossibles à reproduire par une combinaison additive des trois couleurs de base RVB. Par contre, il est toujours possible d'obtenir une équivalence en additionnant à la couleur X le projecteur bleu, et en comparant le résultat obtenu avec le mélange du rouge et du vert.

2.4.7. Représentation spatiale

Si on porte sur trois axes orthogonaux les intensités respectives de Rouge, Vert et Bleu, toute couleur X peut être représentée dans cet espace par un vecteur d'origine O et de longueur proportionnelle à son intensité. En traçant tous les vecteurs correspondant aux couleurs monochromatiques du spectre compris entre 380 et 780 nanomètres, on obtient l'enveloppe des couleurs dont la pureté vaut 1.

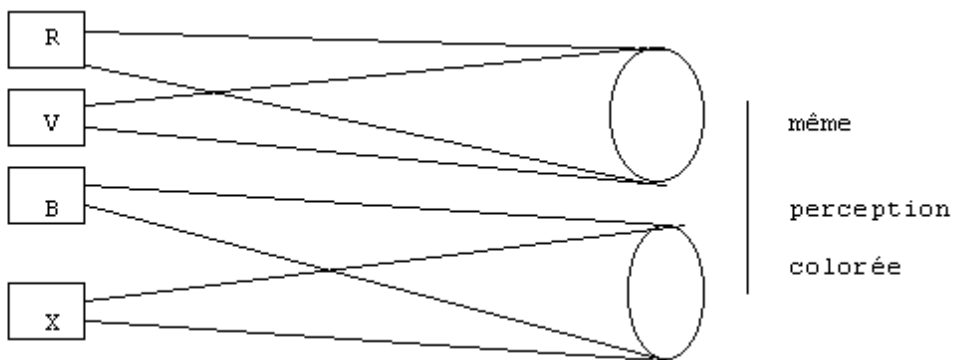


Cet ensemble décrit une surface conique convexe ouverte.

2.4.8. Représentation sur un plan

Pour la commodité des études, il est intéressant de projeter ce volume sur un plan. En projetant le vecteur d'une couleur X sur un plan de luminosité constante, on obtient un point de coordonnées (x,y) dont la position dans le plan est indépendante de la luminosité. L'axe des Y représente l'intensité du vert, l'axe des X l'intensité du rouge. Chaque couleur peut alors être définie par 3 nouvelles coordonnées :

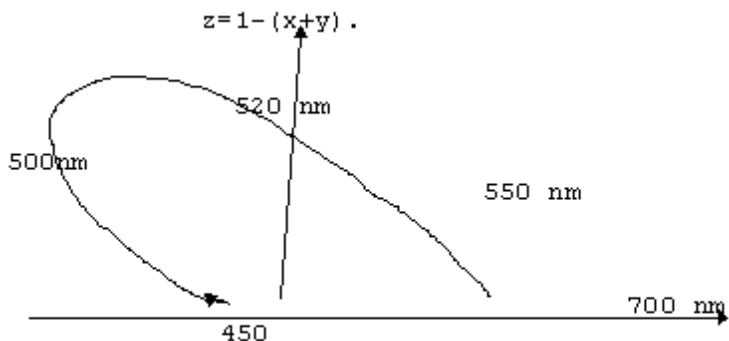
- la position d'un point dans le plan définit la chromaticité, coordonnées x et y.
- la valeur effective de sa luminosité Y pour définir son intensité.



On admet alors que la quantification trichrome de certaines couleurs très pures nécessite une composante négative. Dans ce cas la formule est la suivante :

$$r(R)+v(V)=X+b(B) \text{ soit } X= r(R)+v(V)-b(B)$$

La coordonnée z correspondant à la position par rapport à l'axe des bleus est obtenue par différence :



Cette technique de mesure des couleurs est la base des systèmes colorimétriques. Elle donne une représentation mathématique de la couleur utilisable pour les applications.

3. Différents systèmes colorimétriques utilisés en informatique

Depuis 1931, la Commission Internationale de l'éclairage publie des recommandations concernant, entre autres, les systèmes mathématiques de coordonnées trichromatiques à utiliser dans les domaines industriels et scientifiques. Sans prétendre à l'exhaustivité, voici quelques espaces colorimétriques utilisés par des matériels informatiques pour lire, afficher, ou imprimer des couleurs.

3.1. Système RVB

Le système RVB (rouge, vert, bleu) (RGB en anglais) défini par la CIE en 1931, utilise les trois couleurs fondamentales de la synthèse additive.

Les trois sources précises recommandées par la CIE sont respectivement à :

- 455,8 nm pour le bleu;
- 546,1 nm pour le vert;
- 700,0 nm pour le rouge.

Le système de coordonnées se rapporte au triangle des couleurs présenté ci-dessus.

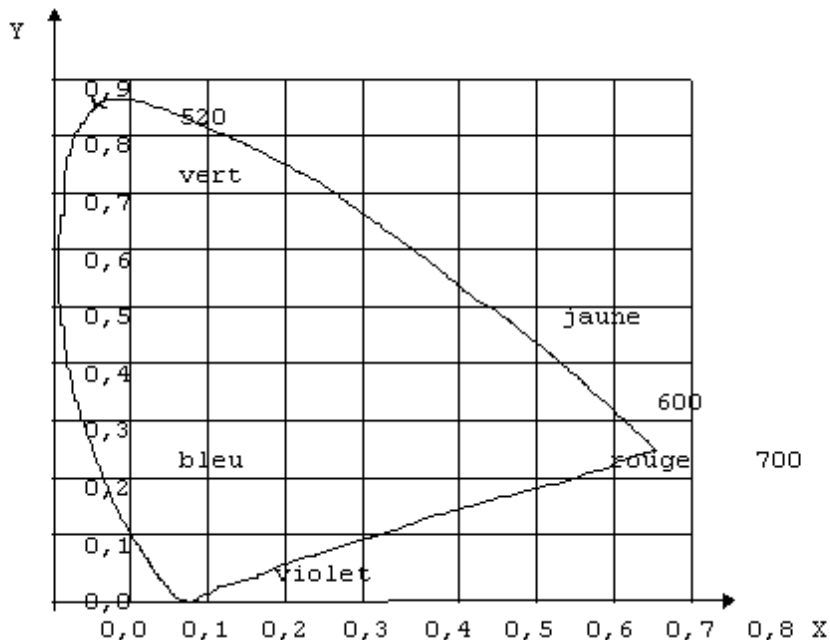
Ce modèle est bien adapté à l'affichage des couleurs sur l'écran.

3.2. Système XYZ (ou Yxy)

Certaines couleurs de teintes bleu-vert ont des coordonnées négatives dans le système RVB et donc ne sont pas constructibles. Pour les atteindre, des manipulations mathématiques ont été réalisées et un nouveau système XYZ a été proposé par la CIE en 1931. Le plan de projection RVB est devenu un plan XYZ tel que, par translation des coordonnées de base :

- toutes les lumières colorées ont des coordonnées trichromatiques positives dans la figure;
- le spectre d'égal énergie, le blanc théorique, est au centre du triangle;
- les nouvelles unités relatives à x et z sont nulles, y donne donc la luminosité.

Cette construction est devenue d'usage universel en colorimétrie.



En 1964, Judd a modifié le système XYZ de 1931 pour tenir compte des champs angulaires de vision à 10°, plus

représentatifs des observations colorées courantes. On note alors des coordonnées x_{10} et y_{10} .

3.3. Système CMJ

Le système CMJ (cyan, magenta, jaune) (cyan yellow magenta) est utilisé en synthèse soustractive. Il est utilisé pour la sortie en couleur de documents sur papier. Les coordonnées d'une couleur sont les proportions de cyan, magenta et jaune. Pour obtenir un noir parfait plus facilement que par superposition des couleurs magenta, cyan et jaune on utilise souvent le noir sous la forme d'une encre spéciale. On parle alors de quadrichromie et de système CMJN (ou CYMB en anglais). Les imprimantes à jet d'encre ont trois cartouches de couleur ou quatre cartouches, trois de couleur et une noire.

3.4. Système HLS ou Y, Db, Dr

Le système HLS, «hue, lighthness, saturation», consiste à définir une lumière par sa teinte (hue), la luminosité ou l'intensité de la lumière (lightness) et la saturation. Ce qui revient à combiner chrominance et luminance. Ce modèle a été imaginé pour coder les signaux vidéo couleur. Lors du lancement des émissions de télévision couleur, il fallait pouvoir recevoir des émissions en couleur sur des équipements noir et blanc. Les trois signaux primaires R, V et B fournis par la caméra vidéo couleur sont alors combinés par une opération appelée *matriçage des couleurs*, selon les trois formules suivantes :

$$Y = 0,30R + 0,59V + 0,11 B$$

$$Db = B - Y$$

$$Dr = R - Y$$

Les coefficients ont été déterminés à partir de la sensibilité spectrale de l'œil humain. Pour le blanc, le noir ou le gris on a $R = V = B = Y$. Le signal Y est alors équivalent à celui fourni par une caméra noir et blanc filmant la même scène.

3.5. Systèmes LAB et LUV

En 1942, *Mac Adam* a démontré par ses travaux que le système Yxy trouve sa limite lorsqu'on veut évaluer la différence entre deux couleurs. Pour cela, il a fait des observations afin d'établir la relation entre la distinction visuelle et l'écart chromatique. À partir d'une couleur donnée, il porte dans le diagramme de chromaticité du système XYZ, les couleurs que l'œil de l'observateur moyen parvient à distinguer comme voisines. L'ensemble des couleurs voisines décrit des ellipses dont la taille et l'orientation sont différentes selon leurs positions dans le diagramme de chromaticité.

En 1976, la CIE a agréé deux nouvelles formules, le système LUV créé par *Wyszecki* et le système LAB d'après les travaux d'*Adams* et de *Nickerson*.

Dans ces systèmes, la luminosité Y du modèle XYZ est transformée mathématiquement afin de présenter une échelle homogène avec la sensibilité de l'œil. Des fonctions complexes sont appliquées aux coordonnées x et y, afin d'obtenir un espace uniforme dans les trois dimensions.

En 1994, la CIE a proposé des améliorations à la formule CIE LAB par une pondération de la contribution des écarts de clarté, de chromaticité et de teinte dans l'écart de couleur total.

Robertson a comparé systématiquement les deux espaces CIELAB et CIELUV. Pour l'uniformité des représentations, aucun système n'a de mérite particulier.

4. Technologie de la couleur en informatique

Pour pouvoir manipuler l'image comme une donnée informatique, il faut la *numériser*. Pour cela on fait de l'image une mosaïque de points : les *pixels* (de l'anglais picture element). À ces points sont attachés des attributs numériques. Une fois mise en mémoire sur un support, l'image codée doit être décodée pour reformer l'image de départ sur un écran ou sur une imprimante.

Nous examinerons :

- les techniques principales de saisie, de mémorisation, d'affichage, de transmission et d'impression des couleurs;
- les divers systèmes colorimétriques utilisés;
- les inconvénients dus à cette hétérogénéité et les remèdes employés.

4.1. Techniques d'acquisition

La saisie d'une image couleur sur un ordinateur, se fait essentiellement par une caméra vidéo (image animée), un scanner ou un appareil photographique (image fixe). Ils transforment l'image en signal électrique analogique. La carte de numérisation transforme le signal analogique en un signal numérique.

4.1.1. Les capteurs à CCD

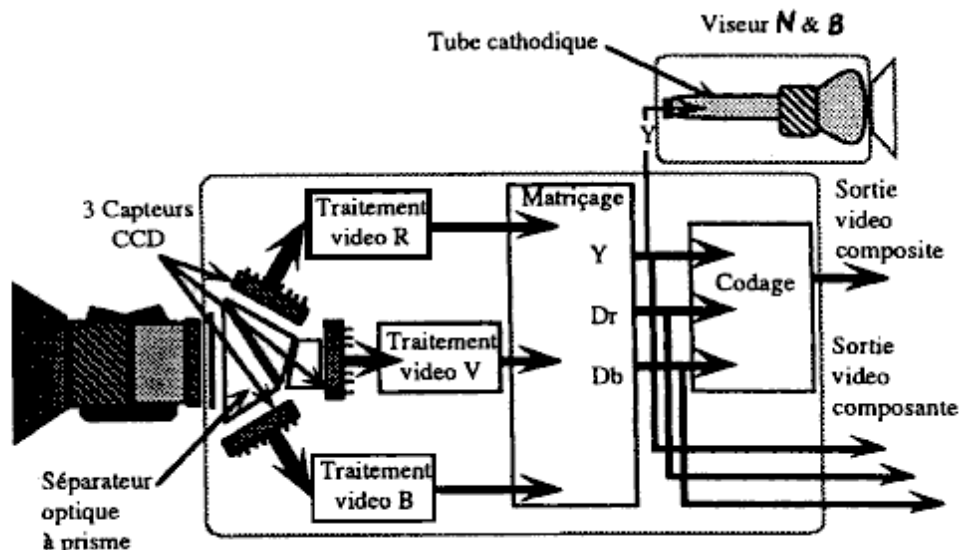
Dans tous les cas, on utilise une cellule optoélectronique munie de récepteurs photo-sensibles dits capteurs à transfert de charge, connus sous le nom de CCD pour «charge coupled device». Ils transforment la lumière reçue en tensions électriques proportionnelles à l'intensité de lumière.

4.1.2. Caméra vidéo couleur

Les *CCD matriciels* sont employés sur les caméras.

Ils ont la forme habituelle d'un circuit intégré à la surface duquel une fenêtre active est dessinée. La surface de cette fenêtre est découpée en cellules élémentaires ou *pixels*. Chacun de ces pixels comprend une cellule photosensible, dont le niveau de charges électriques dépend de la lumière reçue et une cellule de mémoire associée. Les cellules de mémoire sont reliées pour constituer des registres à décalage par transfert des charges. Les charges issues des cellules sensibles sont amenées vers un étage de sortie, en respectant les normes de balayage télévision. À la fin de l'exploration de chaque image, toutes les charges électriques produites par l'impact de lumière sont transférées vers les cellules de mémoire. Ces dernières sont vidées par transfert de charges, au fur et à mesure du balayage vidéo, vers un registre qui met en forme le signal vidéo.

Les caméras vidéo couleur sont basées sur une *analyse trichrome* qui décompose l'image en 3 couleurs primaires.



Principe d'une caméra vidéo couleur tri-CCD.

Dans les *caméras couleurs tri-CCD*, le faisceau lumineux capté par l'objectif est divisé en trois faisceaux distincts par un jeu de prismes. Deux des faces de ces prismes agissent comme des filtres dichroïques et le divisent ainsi en trois faisceaux rouge, vert et bleu vers trois capteurs CCD, qui font l'analyse trichrome. Le positionnement mécanique des trois analyseurs CCD est très précis, car à un moment donné, il faut que l'analyse du pixel de chacun des CCD corresponde au même point lumineux de l'image.

Pour obtenir un meilleur recouvrement de l'analyse de l'image et compenser la faible surface des pixels des analyseurs, le CCD vert est décalé selon l'axe horizontal de la valeur d'un demi pixel. La précision de montage est inférieure au 1/100 de mm. Le coût élevé réserve cette technologie aux secteurs institutionnel et professionnel.

Dans les *caméras couleurmono-CCD*, le CCD unique est doté d'un filtre à fines raies verticales alternatives, rouges, vertes et bleues. Les trois signaux rouge, vert et bleu sont séparés par indexation électrique au niveau du CCD. Afin de limiter la perte de définition, une corrélation dans le sens vertical et horizontal est faite pour à reconstituer les zones d'information manquantes. La définition et la sensibilité des caméras mono-CCD est moins élevée que celle des tri-CCD. Leur simplicité réduit leur coût et les destine au grand public. Dans toute caméra vidéo couleur, en plus d'un ou trois capteurs CCD, se trouvent d'autres dispositifs spécifiques pour le traitement des couleurs:

- Un *système de filtres rotatifs* est monté entre l'objectif et le bloc optique de séparation pour adapter les caractéristiques colorimétriques des CCD à la température de couleur de la lumière éclairant la scène filmée. Le filtre est choisi selon le type et les conditions d'éclairage (lumière du jour, éclairage artificiel, ...).
- Un *réglage électronique de balance de blanc* affine le choix du filtre. Il s'agit d'un dispositif électronique de correction colorimétrique, automatique, semi-automatique ou manuel). Le principe de ce réglage consiste à filmer une surface blanche et à modifier le gain des voies rouge et bleue de sorte à rendre identique les trois valeurs R, V et B, puisque la surface blanche fournit une lumière égale dans toutes les composantes.
- Divers *circuits de traitement correcteurs et adaptateurs* rectifient les signaux de manière à tenir compte des courbes de transferts des CCD et des tubes cathodiques servant à l'affichage: circuits d'écrêtage, corrections de contour, réglage de niveau de noir, correction de pente dans les blancs, etc.
- Un *codeur* transforme les signaux RVB obtenus en sortie de ces circuits de traitement, en signaux PAL, NTSC, Y/C ou S-vidéo. Ces signaux vidéo sont dérivés du système Y, Db, Dr avec codage sur un ou trois circuits.

4.1.3. Scanners ou numériseurs

Les scanners couleurs sont des outils de conversion d'images fixes noir et blanc ou couleur en images numériques. Les plus courants sont les *scanners à plat*. On présente le document à numériser sur une vitre. Une barrette linéaire de CCD balaye la surface du document à convertir. Chaque cellule de la barrette convertit le niveau lumineux du point exploré en niveau électrique. La prise de vue se fait donc par un balayage ligne par ligne, d'où l'expression anglo-saxonne «to scan»: parcourir, scruter, balayer. Les *scanners à main*, utilisent la même technique.

Pour filmer des objets en trois dimensions, il existe des *pseudo-scanners* constitués d'une caméra vidéo montée sur banc-titre.

La qualité de la lecture d'une couleur sera liée à la *définition* ou *résolution* du scanner, qui est donnée en *ppi* «points per inch» ou *dpi* «dots per inch», points ou pixels par pouce. Les valeurs courantes sont de 300 à 2500 dpi. Une résolution élevée conduit à des fichiers énormes. La *profondeur de quantification* ou *profondeur d'analyse* peut varier de 1 bit (noir ou blanc) à 30 bits en RVB (pour 1,07 milliard de teintes).

Un scanner est fourni avec une carte d'interface ou un raccordement à la voie parallèle, ainsi qu'avec des pilotes ou un logiciel pour convertir l'image acquise dans des formats graphiques habituels et éventuellement reconnaître les caractères.

Technique des scanners.

Ils utilisent deux types de capteurs, les tubes photomultiplicateurs (PMT pour «photomultiplier tube») et les CCD.

Dans un PMT, un tube à vide reçoit la lumière, la transforme en tension électrique et amplifie le signal de sortie. La réception d'un photon peut être transformée en l'équivalent de plusieurs centaines de photons. Les fournisseurs habituels sont Hell, Crosfield et Dainippon screen. Leur caractéristiques sont la grande résolution et la grande largeur de bande. À notre connaissance, Optronics colorGetters fournit des scanners de bureau.

Les scanners de bureau, le plus grand marché en nombre, utilisent la technique CCD déjà décrite.

4.1.4. Cartes de numérisation vidéo

Ces cartes transforment le signal vidéo analogique en fichier informatique.

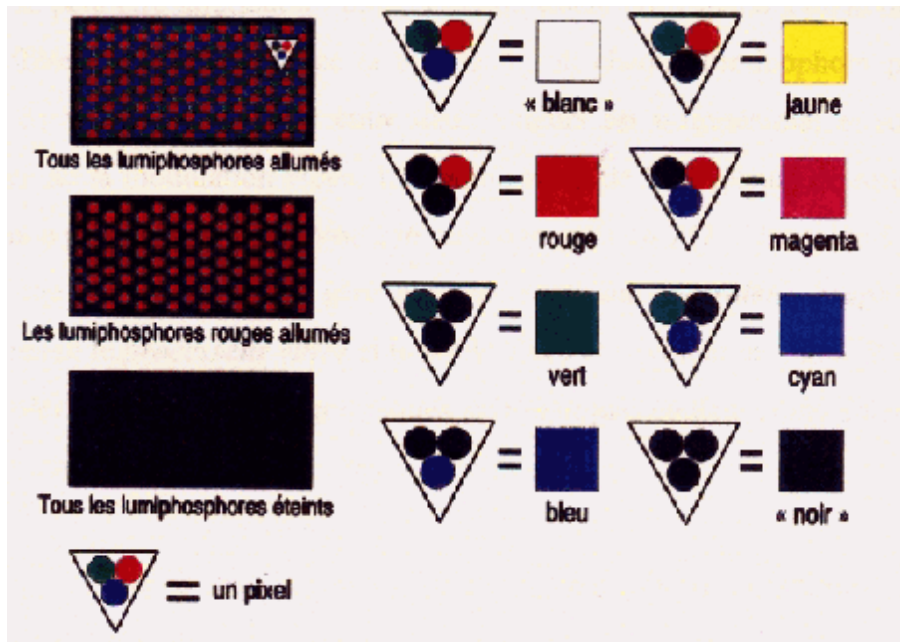
Les caractéristiques d'une carte de numérisation sont:

- le type de bus;
- la norme de couleur et le standard de balayage des signaux vidéo d'entrée (RVB, PAL, SECAM, NTSC, S-Vidéo);
- le format de l'image numérisée en pixels, le plus courant est le format 640x480. Le format 720x576 de la vidéo numérique est annoncé;
- la profondeur de quantification qui influe sur le nombre de couleurs reproduites;
- la durée de numérisation;
- d'autres fonctions implantées sur la carte comme affichage, conversion vidéo, etc.
- les formats graphiques dans lesquels sont enregistrés les images;
- les logiciels de numérisation et des pilotes de commande et de réglage.

4.2. Techniques d'affichage sur écran

4.2.1. Comparaison entre écrans informatique et télévision

Les écrans graphiques d'ordinateur sont improprement appelés moniteurs vidéo. Nous allons voir en quoi ils ressemblent et en quoi ils diffèrent des écrans de télévision couleur. Certes, ils utilisent comme les écrans de télévision, la *synthèse additive*. L'affichage sur écran d'ordinateur, comme sur écran de télévision, repose sur le système RVB.



Chaque pixel d'un moniteur couleur est constitué de trois phosphores: rouge, vert, bleu. Chacun est excité par un pinceau électronique spécifique. Il émet une lueur rouge, verte ou bleue. Chaque phosphore émet sa propre intensité. Par illumination sélective, les trois phosphores créent la base des couleurs : noir, rouge, vert, bleu cyan, magenta, jaune et blanc. Le noir est obtenu par les trois phosphores éteints et le blanc en les ayant tous allumés. La finesse des points RVB sur l'écran est telle que l'œil ne peut les dissocier. La perception correspond au mélange additif de ces 3 couleurs primaires.

La grande différence avec la télévision analogique réside dans le fait que chaque luminophore ne peut émettre

qu'un nombre défini et limité d'intensités. Par exemple, si chaque phosphore peut afficher 256 intensités différentes à l'écran de l'ordinateur (variant entre 0 et 255), il existe 256 valeurs de rouge, 256 valeurs de vert et 256 valeurs de bleu. Le rouge correspondant à la valeur 0 est en réalité du noir, le rouge correspondant à la valeur 1 ou 2 est un rouge très sombre, presque noir. Celui qui correspond à la valeur 255 est un rouge très saturé. L'écart entre deux valeurs ne peut être inférieur à l'unité. La combinaison de 256 valeurs de rouge, de vert et de bleu donne par synthèse trichrome 2563 couleurs, soit 16 777 316 couleurs. Dans un récepteur de télévision analogique, chaque luminophore prend ses valeurs selon les variations continues de la modulation vidéo. Leur nombre est limité par le bruit. La *carte ou contrôleur graphique* comprend le *processeur vidéo* et la *RAM vidéo*. Elle envoie un signal RVB à l'écran.

À la sortie des premiers écrans graphiques informatiques et pour simplifier l'électronique des cartes, les constructeurs avaient choisi de fournir les signaux RVB couleur sous forme de signaux TTL de 0 à 5V, forme habituelle des signaux logiques à l'intérieur des micro-ordinateurs. Électroniquement simple, cette solution présentait deux inconvénients. D'une part les signaux ne pouvant prendre que 2 états électriques (0V ou 5V), chaque signal de couleur RVB fonctionnait en tout ou rien ce qui donnait 32, soit 8 couleurs au total. Avec des niveaux de demi intensité, on passait à 64 couleurs.

L'autre inconvénient se situait dans le transport des signaux qui ne comprennent que des fronts montants ou descendants, chargés d'harmoniques de fréquences élevées. Un câble un peu trop long les atténue rapidement et perturbe la reproduction des images couleur.

Pour ces diverses raisons, les constructeurs ont choisi depuis plusieurs années de relier les écrans graphiques aux micro-ordinateurs par des *liaisons analogiques* variant de 0 à 700 mV, correspondant à ceux de la production vidéo. Par contre les fréquences d'affichage (balayage vertical et horizontal) sont différentes de celles des équipements vidéo européens. De plus, pour améliorer la stabilité des images couleurs sur les écrans graphiques d'ordinateur, le balayage des points rouge, vert et bleu n'est pas entrelacé contrairement aux équipements vidéo.

Malgré l'arrivée du multimédia, la vidéo et l'informatique restent donc, pour l'instant, encore deux mondes à part.

4.2.2. Écrans à cristaux liquides

On retrouve cette technique dans les micro-ordinateurs portables et dans les tablettes LCD pour rétroprojecteurs.

L'écran est constitué d'une *matrice de cellules LCD (Liquid Cristal Display)* dont le niveau individuel de transparence varie avec la tension électrique appliquée. On peut ainsi reconstituer l'image pixel par pixel selon les normes habituelles de balayage vidéo. Les tablettes LCD couleur comprennent trois réseaux de cellules LCD colorées superposés en RVB.

4.2.3. Caractéristiques techniques d'un écran

La *résolution graphique* ou *définition* de l'écran est le format de l'image affichée. La moyenne résolution est de 800x600 pixels (écrans VGA). La haute résolution 1024 x 768 existe en superVGA PC et sur les grands écrans des Macintosh. Pour les images animées, la résolution employée en vidéo professionnelle est de 768 x 576 points.

La résolution doit être compatible avec la taille du tube cathodique mesurée en pouces (de 12" à 21"). Les hautes résolutions ne s'expriment bien qu'avec de grands écrans, 16 pouces et au delà.

Pour conserver une échelle constante aux images reproduites, les dimensions de l'image affichée plein écran (correspondant à celle du plan mémoire vidéo) croissent avec la taille de l'écran. Les formats habituels sont indiqués dans le tableau suivant.

Taille du tube en pouces	Largeur de l'image en pixels	Hauteur de l'image en pixels	Taille du fichier à 256 couleurs en ko	Taille du fichier à 16 millions de couleurs en ko
13 ou 14	640	480	300	900

15	870	640	544	1632
16 ou 17	832	624	507	1521
19	1021	768	768	2304
21	1152	870	969	2936

Formats d'écran avec taille de l'image plein écran

On constate la croissance géométrique des tailles de fichiers et surtout le caractère irréaliste d'applications utilisant des dimensions d'images trop importantes.

L'autre paramètre indiqué par les constructeurs est *le pas des luminophores* ou «*pitch*». Il définit l'espacement des luminophores sur la face intérieure du tube cathodique. Plus le pas est faible, plus l'image est fine.

Le nombre de couleurs accessibles est un facteur encore plus important d'un écran graphique. L'ensemble des couleurs affichables résulte de combinaisons de couleurs de base dans la *table de couleurs (LUT ou look-up table* en anglais). Pour modifier la palette des couleurs affichées ou affichables, il suffit de modifier la table des couleurs. C'est ce qu'on fait lorsqu'on utilise l'icône couleurs du panneau de configuration de Windows.

4.3. Techniques d'impression

Les imprimantes sont basées sur le système colorimétrique CMJ ou sa variante CMJN. Elles utilisent *la synthèse soustractive*.

Selon le procédé, l'image est reconstituée sur le support :

- par dépôt de poudre et thermofusion, pour les imprimantes laser;
- par projection de gouttelettes d'encre, pour les imprimantes à jet d'encre;
- par décalque à chaud d'une pellicule d'encre, pour les imprimantes thermiques.

Dans les descriptions qui suivent on parlera d'encre, mais le phénomène décrit est le même pour la poudre.

Les imprimantes courantes contiennent les encres cyan, jaune et magenta. En superposant ces trois encres à égalité, un point a l'une des huit couleurs : noir, rouge, vert, bleu, cyan, magenta, jaune et blanc. Si aucune encre n'est déposée, on obtient du blanc, la couleur du papier. Le noir est la superposition des trois encres.

Mais le plus souvent on utilise *la quadrichromie* et le système *CJMN*. Dans ce cas on utilise une quatrième encre, noire pure, afin d'obtenir un noir plus homogène que le noir composé et certains tons foncés. Cette encre noire est aussi utilisée seule pour des documents rédigés en noir et blanc.

Récemment, Hewlett-Packard et Canon ont choisi d'imprimer avec six encres de couleur pour obtenir davantage de nuances.

L'imprimante peut donc produire huit couleurs de base. Un gestionnaire d'imprimante peut employer différentes méthodes pour mélanger ou disposer correctement ces huit couleurs, afin d'obtenir la nuance désirée. Nous allons examiner deux méthodes couramment employées.

4.3.1. Mode motif

Le mode Motif produit plus de nuances car il définit une matrice de pixels plus grande appelée *motif de simili*. Il s'agit d'une matrice de 8x8 points. À l'impression, ce motif apparaît sur le papier comme une nuance de couleur bien qu'il ne soit qu'une combinaison de points de la base des couleurs.

Pour remplir une zone d'une nuance particulière, les motifs en simili sont disposés en rangées et colonnes comme dans un tableau. Par le mélange des huit couleurs, le gestionnaire d'imprimante peut pratiquement représenter sur le papier des millions de nuances. Ces motifs ont l'inconvénient de pouvoir être visibles.

4.3.2. Mode nuage

Sous Microsoft Windows et Macintosh, on utilise souvent le mode nuage. Cette méthode utilise la même technique que la méthode motif, à l'exception du placement des jets d'encre. Ils sont disposés au hasard et suppriment l'inconvénient des motifs.

4.4 Mémorisation des données des images en couleur

4.4.1. Images fixes

Le besoin de modéliser et de stocker des images fixes provient de domaines d'applications nombreux et variés : arts graphiques, PAO, CAO, PréAO, Bureautique, gestion documentaire, cartographie, télédétection, imagerie médicale.

Les besoins sont donc très divers, tant sur les données elles-mêmes que sur les traitements. Néanmoins deux grandes classes de format et de structures de données sont couramment admis: les *formats vectoriels* pour les *données graphiques* et les *formats matriciels* (ou *bitmap* ou encore *raster*) pour les données images.

4.4.2. Données d'images et représentation matricielle

Du point de vue photographique, une image (analogique) est une fonction 2D de l'intensité de la lumière avec des niveaux de discrétisation.

Du point de vue informatique, une image (numérique) est constituée d'une matrice de pixels.

Les données analogiques sont enregistrées sur des supports appropriés de type vidéodisque, contenant environ 54000 images).

Pour les données numériques, l'enregistrement des valeurs de chaque pixel, avec ou sans compactage ou compression, est fait sur support informatique: disque, bande, disque optique numérique ou CD-ROM.

La représentation matricielle est adaptée aux périphériques travaillant en mode ligne ainsi que pour le stockage et le traitement d'images complexes. Les conversions sont relativement faciles entre formats matriciels, sauf pour les formats compressés.

La représentation matricielle présente des inconvénients:

- Il s'agit d'un *codage pauvre de l'information*, dans la mesure où la sémantique n'est pas prise en compte, et où l'on ne distingue pas les objets qui composent l'image.
- C'est une représentation qui n'est pas adaptée à l'échelle variable : le zoom sur une image s'accompagne de dégradation visuelle, effet de marche d'escalier. D'autre part, l'opération inverse (effet de type grand angle) oblige à effectuer des regroupements de pixels qui ne sont pas toujours simples à réaliser si on veut conserver une bonne qualité de l'image et une bonne restitution des couleurs.
- La représentation matricielle n'est pas compacte, car elle conserve tous les points de l'image, sauf si l'on utilise des méthodes de compression qui entraînent une perte de temps, et éventuellement une perte d'information. Les traitements d'images sont longs, sauf si on dispose de matériel spécialisé.

4.4.2.1. Formats matriciels de stockage et d'échange des données

BMP utilisé par Windows, gère 24 bits de couleur.

EPS «encapsulated postscript file» d'Adobe, à écriture seulement créé par Adobe pour que des images puissent être insérées dans un traitement de texte et être imprimées en postscript. Ce fichier ne peut pas être lu ni s'afficher sur un écran.

Kodak Photo CD créé par Eastman Kodak en 1992 est plus professionnel que grand public. Sa résolution s'approche

du film 35 mm. Il est géré par Unix, Windows NT, Mac, PC... Il avait été conçu pour le CD contenant au plus 100 photos de qualité 35 mm ou 800 photos de qualité TV. Un logiciel permet d'adapter la résolution et les couleurs en fonction des besoins. En 1995, deux catégories de format étaient autorisées: le format maître et le format de distribution supportant 6 niveaux de résolution de 512 x 768 jusqu'à 4096 x 6144 en vraie couleur. Il est à noter que seul un lecteur CD-ROM avec la norme XA et multisession peut décoder les différents formats du Kodak Photo CD.

GIF CompuServe «graphics interchange format» est très répandu. L'image est analysée ligne par ligne. Le codage est sur 8 bits (256 couleurs), il peut stocker conjointement trois plans images (rouge, vert, bleu), ce qui simule un codage sur 24 bits (16 millions de couleurs). Les algorithmes de compression employés sont de type LZW. Il autorise les échanges entre compatibles PC et Macintosh. Il est le format usuel des images incluses dans des pages html.

PCX appartient à la société ZSoft (logiciel Paintbrush). Il était destiné à la sauvegarde d'écrans en mode CGA. Il a été modifié pour traiter des images 8 bits, mais il est mal adapté aux couleurs sur 16, 24 ou 32 bits. Les algorithmes de compression utilisés sont de type RLC. *PCX* est accepté par de nombreux logiciels pour la récupération de données, il est peu utilisé en création.

TIFF «tagged image file format» a été créé par Aldus, Microsoft et NeXT, pour être un standard de codage des images scannées. Il est répandu et utilisé pour les échanges. Presque tous les logiciels l'acceptent en entrée et sortie, sur PC comme sur Macintosh. Il code tous les types d'images, mais est très complexe à gérer. Les compressions sont RLC ou LZW. Il s'adapte et évolue pour prendre en compte les données multimédias (codage d'objets).

Il y avait en 1998 plus de 160 formats en cours d'utilisation. Le lecteur curieux chargera le logiciel partagé PaintShop Pro, il y relèvera les formats acceptés en entrée et sortie. La diversité de ces formats rend encore plus complexe l'échange et la restitution fidèle d'images couleurs entre différents systèmes.

4.4.2.2. Données graphiques et représentation vectorielle

Par opposition aux données images les données graphiques sont généralement considérés comme des données riches mais complexes.

On distingue trois types de données selon qu'elles sont représentées par :

- des points isolés pour modéliser des objets ponctuels;
- des lignes isolées pour des objets linéaires;
- des surfaces isolées ou par des mosaïques pour des objets surfaciques.

Les données graphiques ont des représentations vectorielles.

On stocke la description géométrique de l'objet et ses attributs comme la couleur. Par exemple, un cercle rouge ne sera pas stocké par représentation matricielle de l'image en pixels, mais par un objet cercle de rayon r et d'attribut couleur rouge. Cependant, les données graphiques ne sont pas adaptées aux images complexes qui contiennent un grand nombre d'objets de petites tailles.

De plus la complexité des formats vectoriels rend difficiles les opérations de conversion entre formats.

4.4.3. Images animées

L'impression de mouvement est donnée par une succession d'images fixes reproduites à une vitesse suffisamment élevée. Chaque image devient un élément indépendant décrit dans sa totalité, le temps étant un paramètre fixe. On se base sur les mêmes formats que ceux définis pour les images fixes et décrits précédemment.

4.5. Techniques de compression

Un des problèmes majeurs du multimédia concerne la vidéo couleur. Les débits de vidéo numérique sont de l'ordre de 200 Mbits/s. Même en se limitant à une qualité et à une définition moindre, une seule image en mode VGA étendu occupe en effet 307 Ko (640x480 points en 256 couleurs, codées sur 8 bits).

Pour afficher une animation à la cadence de 25 images par seconde, il faut un débit de 7,6 Mo/s (25X307 ko/s). Or le débit des lecteurs de CD-ROM est de l'ordre de 13,3 Mbits/s en octuple vitesse soit 1,66 Mo/s. Un facteur de compression de l'ordre de 5 est alors indispensable.

Avec des images de qualité photographique, 16 millions de couleurs codées avec 24 bits, il faudrait un débit 3 fois plus grand et par conséquent un taux de compression de l'ordre de 15.

La difficulté est encore plus grande pour la transmission sur des réseaux même en utilisant les 64 kb/s de Numéris.

Enfin, outre les débits, la capacité de stockage d'un CD-ROM est de 650 Mo. Si les images n'étaient pas compressées on ne pourrait stocker qu'environ une minute et demie de vidéo sur ce support.

Il y a donc nécessité de réduire le volume occupé par l'information par des méthodes de *compactage*, sans perte, ou de *compression*, destructive. Des normes de compression ont été définies afin d'éviter la multiplication des techniques et de limiter les risques d'incompatibilité. Il s'agit essentiellement des normes *JPEG* (pour les images fixes) et *MPEG* (pour les images animées).

4.5.1. Norme JPEG pour compression d'images fixes

Cette normalisation a été faite par un groupe d'experts, au sein du comité JPEG (Joint Photographic Expert Group), créé à l'initiative de l'ISO, du CCITT et de nombreux constructeurs.

Le principe du codage est le suivant :

1. l'image RVB est transcodée en Y Db Dr;
2. elle est ensuite divisée en blocs de 8x8 pixels ou imquettes;
3. pour chaque bloc de 64 pixels, on applique la transformation en cosinus discret ou DCT (discrete cosine transform). On passe ainsi d'une représentation spatiale à une représentation fréquentielle avec une composante continue donnant le niveau moyen du bloc;
4. les fréquences les plus élevées sont supprimées, car elles ne sont pas perçues par l'œil;
5. un codage de type Huffman est appliqué.

Des taux de compression de 20 à 25 sont possibles sans altération visible de l'image. Des informations de détail très fines sont néanmoins perdues.

4.5.2 Quelques formats d'images

4.5.3. Norme MPEG de compression d'images animées

Son principe est de coder les différences entre images successives au lieu des images elles-mêmes. Le groupe de travail «moving pictures expert group», de l'ISO, a créé les normes MPEG-1, MPEG-2, MPEG-3 et travaille actuellement à la norme MPEG-4.

MPEG-1 utilise le même algorithme que la norme JPEG. La taille d'image est basse 352x240 ou 352x288 pour un débit de 1,2 Mbits/s. La couleur est analysée par blocs de 8x8 pixels. La prédiction des mouvements s'effectue à partir de l'intensité sur des blocs de 16x16. Chaque bloc 16x16 est comparé avec son homologue dans une image passée ou future, pour calculer et quantifier les différences par un coefficient.

MPEG-2 étend MPEG-1 à des résolutions plus élevées, 352x576 ou 720x576 pour un débit de 20 Mbits/s au plus. Elle est faite par des cartes de compression et décompression.

MPEG-3 vise des débits de 20 à 40 Mbits/s. Elle est maintenant englobée dans la norme MPEG-2.

MPEG-4 sera applicable à des canaux à débits faibles, entre 4800 et 64000 bits par seconde. Elle innovera avec un mode d'encodage basé sur une analyse des objets contenus dans l'image (contour, texture ...).

4.6. Harmonisation des couleurs par les CMS

Nous avons vu que les techniques de lecture, de conservation, d'affichage et d'impression des couleurs en informatique sont fondées sur des systèmes colorimétriques différents. L'obligation de compression mène à une perte d'information.

Les conséquences sont de peu d'importance pour l'impression d'un graphique. Par contre, en infographie, il est indispensable d'obtenir une représentation fidèle des couleurs à chaque étape du traitement de l'image.

Les espaces colorimétriques RVB et CMJN possèdent de nombreuses divergences. Certaines teintes présentes dans l'un ne le sont plus dans l'autre. Autrement dit, toutes les nuances saisies par un scanner ne pourront pas être reproduites par une imprimante numérique, ni même une presse offset. Or toutes les images numérisées en mode RVB devront finalement être converties dans l'espace colorimétrique CMJN afin d'être imprimées.

De plus, dans le même espace colorimétrique, deux moniteurs RVB de marques différentes auront des gammes de couleurs ou *gamuts* différentes, donnant à voir des couleurs différentes et faussant la retouche à l'écran. Un bon scanner saisira plus de nuances qu'un modèle de piètre qualité. De même, deux imprimantes peuvent rendre différemment un même fichier image.

Le calibrage de la chaîne infographique peut tourner au casse-tête.

Pour résoudre ce problème, il fallait donc trouver un système capable de gommer tous ces facteurs d'écarts. Il s'agit de prendre en compte les espaces colorimétriques propres à chaque périphérique présents dans une chaîne de fabrication afin de les harmoniser. Ces systèmes logiciels sont appelés CMS (*système de gestion des couleurs ou colour management system*).

Plusieurs CMS étaient disponibles sur le marché en 1997, Eficolor d'Electronic for imaging, que l'on trouve dans Quark Xpress, FotoTune d'Agfa livré en version allégée avec tous les scanners de ce constructeur, Precision de Kodak, présent dans Pagemaker 6.0 d'Adobe et Colorsync d'Apple 2.0 fourni avec le système 7.5.x.

La première tâche d'un CMS va être de *caractériser un périphérique* donné non plus dans son gamut dépendant (RVB ou CMJN), mais *dans un espace colorimétrique indépendant de tout matériel*. C'est en général sur le modèle CIE Lab que se fondent les CMS pour situer le gamut des périphériques.

Chaque périphérique est livré avec un fichier, nommé *profil*, qui contient son gamut propre dans le modèle colorimétrique CIE. Le CMS analyse et compare les profils du scanner, du moniteur et de l'imprimante (voire même d'une seconde imprimante pour la sortie finale, la première étant souvent consacrée au bon à tirer). Dès lors le CMS connaît *l'espace colorimétrique commun* à tous ces appareils, c'est-à-dire les couleurs qui seront numérisées, visualisées et imprimées sans qu'aucune retouche ne soit nécessaire. Chaque CMS est fourni avec une bibliothèque de profils couvrant les matériels les plus couramment utilisés. Un professionnel de l'image étalonne son scanner et son imprimante et crée ses profils personnalisés.

Garantir l'invariance des couleurs de la numérisation d'une photographie à son impression en quadrichromie est impossible aujourd'hui. Bien entendu, les périphériques progressent de mois en mois pour étendre leur espace colorimétrique. Toutefois, il y a des limites qui ne sont pas près d'être franchies. L'harmonisation des divers maillons de la chaîne et la maîtrise de leurs imperfections sera la réponse efficace au problème de la portabilité de la couleur.

5. BIBLIOGRAPHIE

OUVRAGES

- Y. Dordet, **La colorimétrie, principes et applications**, Eyrolles, 1990.
- P. Boursier, P-A. Taufour, **La technologie multimédia**, Hermès, 1994.
- D. Legrand, **La couleur imprimée - mode d'emploi**, Trait d'union graphique, 1990.
- F. Louguet, **Synthèse d'images sur micro ordinateur**, Techniques, Matériels, Logiciels, Dunod Tech, 1992.
- C. Chaillou, **Architecture des systèmes pour la synthèse d'images**, Dunod Informatique, 1992.
- J. Ségura, **Du scanner aux images numériques**, Nathan, 1989.

- E. Couchot, Images - De l'optique au numérique*, Hermès, 1988.
M. Bret, Images de synthèse, Dunod, 1988.
B. Péroche, J. Argence, D. Michelucci et D. Ghazanfarpour, La synthèse des images, Hermès, 1988.
R. Sève, Physique de la couleur, de l'apparence colorée à la technique colorimétrique, Masson, 1996.

BROCHURES TECHNIQUES

Guide d'utilisation de la couleur (Hewlett Packard).

Manuel technique du moniteur couleur P50 (Compaq).

On trouvera une documentation très complète sur le format libre PNG (portable network graphics) à :

<ftp://ftp.uu.net/graphics/png/documents/pngextensions>

REVUES

Le Monde Informatique - 20/09/96 - Compression et décompression vidéo.

Le Monde Informatique - 01/11/96 - Compresser l'image animée.

Le Monde Informatique - 01/12/96 - Numérique et multimédia.

Réseaux et Télécoms - 01/07/96 - À la poursuite de la vraie couleur.

Sites internet

L'un des plus documentés est <http://www.scantips.com>

On y trouvera de nombreuses références sur :

- les scanners à plat - fabricants et sites;
- les scanners pour films - fabricants et sites;
- les logiciels de traitement d'images;
- la gestion de documents et reconnaissance optique (OCR);
- des informations générales;
- la couleur, la gestion de la couleur (color management systems) et logiciels;
- la photographie;
- les forums.



Conservatoire national des arts et métiers
Architectures des systèmes informatiques
ANNEXE 04
Gestion de la couleur
Année 2002-2003

Conservatoire national des arts et métiers
Architectures des systèmes informatiques
ANNEXE 05
Comparaison des bus WorldFip et Profibus
Année 2002-2003



ou



Définition d'un bus de terrain

Les usines, les bâtiments, les voitures, les avions sont de plus en plus automatisés pour assurer sécurité, fonctionnalité et confort à l'être humain.

Pour ce faire on a besoin de récolter des informations sur l'environnement par l'intermédiaire d'éléments simples tels que des capteurs et de les analyser. On fait traiter ensuite ces données par des outils, automates programmables ou ordinateurs. Une fois les données analysées et traitées, on agit sur l'environnement par l'intermédiaire d'actionneurs si besoin est.

On peut retrouver des éléments récoltant des informations tels que des capteurs de températures, des capteurs de pression, des détecteurs d'incendie, des détecteurs de personnes... et des éléments apportant une action tels que des moteurs pas à pas, des électrovannes, des champs électromagnétiques.

On comprend bien que tous ces éléments aient besoins de communiquer entres eux, dans certains cas comme pour la sécurité, la communication doit se faire de façon simple, rapide et fiable; dans d'autres cas, pour des motifs de coût, sont implémentation doit elle aussi être simple, rapide et fiable, par exemple : Lors de l'assemblage d'une voiture, la connexion des capteurs doit pouvoir se faire par un opérateur sans qu'il ait des connaissances techniques particulières. Ces éléments de par leurs nombres engendrent une multitude de fils nécessaires à leur alimentation et à la transmission de leurs données. Le bus de terrain (FieldBus) prend en compte toutes ces exigences, il a été conçu pour ça. Sa composition peut être décrite en deux parties, une partie matérielle composée d'un média physique sur lequel transitent l'alimentation électrique des éléments définis ci-dessus et leurs transferts de données et une partie logicielle composée d'une multitude de composantes logicielles fournissant le protocole de communication, les échanges d'informations, la supervision, la sécurité, les interfaces et bien d'autres fonctions.

Domaines d'application

Le bus de terrain peut être utilisés pour divers besoins, la sécurité, le confort, l'assistance, l'aide à la décision, la fonctionnalité, etc., l'utilisation de ces besoins peuvent être cumulés suivant les domaines d'application. Les domaines d'application peuvent être un bâtiment, une usine, une voiture, un bateau, un train, une centrale nucléaire, un hôpital.

Domaine d'application : Le bâtiment.

Pour des besoins de sécurité, les bâtiments haut de gamme possèdent des capteurs permettant de détecter la fumer provenant d'un incendie. L'ordinateur relié aux capteurs analyse les informations provenant des capteurs et envoie un ordre de fermeture à un mécanisme de commande des portes automatiques pour cantonner l'incident et un ordre aux actionneurs de commande des buses pour qu'elle propulse le liquide afin d'éteindre l'incendie. Pour des besoins de confort, les bâtiments climatisés possèdent des capteurs de température et des arrivées d'air régulés dans chaque pièce, si la température d'une pièce venait à changer, l'automate envoie un ordre au mécanisme régulant l'arrivée d'air pour obtenir la température voulue.

Domaine d'application : La voiture.

Pour des besoins de sécurité, les constructeurs propose un système de freinage qui ne bloque pas les roues en cas de freinage brusque même sur route glissante. Le système possède des capteurs qui relève la vitesse de rotation des roues et l'envoie à un calculateur embarqué qui analyse et traite l'information et envoie un ordre au mécanisme agissant directement sur les freins.

Origine de WORLDFIP et PROFIBUS

PROFIBUS

Profibus (PROcess FIEld BUS) est le résultat d'un projet lancé en 1989 par le ministère fédéral allemand de la recherche et de la technologie, il a été développé et financé par des entreprises d'automatismes comme Siemens.

WORLDFIP

Worldfip (World Factory Instrumentation Protocol) est une organisation française à but non lucratif fondée en 1988, elle fonctionne grâce au financement des membres de l'organisation, à la mise en place de formations, à l'assistance technique et au support. Un consensus entre les utilisateurs finaux, les institues de recherche et les fabricants est mis en place avant de prendre toute décision visant à faire évoluer le bus. Worldfip est membre de « FieldBus Organisation » et est très actif dans le projet Européen NOAH (Network Oriented Applications Harmonisation). Les buts de Worldfip sont de réduire les coûts de connexion, d'obtenir un maximum de fonctionnalités, d'obtenir un bon débit et d'assurer une intéropérabilité avec le plus grand nombre possible de matériel.

Les bus concurrents

AS-I : (Actuator Sensor Interface) 1993

CAN : Controller Area Network en 1993

Controlnet : 1995

DeviceNet : Milwaukee 1994

Interbus : 1984

LonWorks : Californie 1991

P-Net : Danemark

SDS : Smart Distributed System en 1989

Fonctionnement de WORLDFIP

Termes utilisés :

Variable : Unité d'échange la plus petite entre les stations.

Message : Unité d'échange plus grande et plus explicite que les variables.

Station : Élément du bus pouvant être, un ordinateur, un automate, un capteur ou un actionneur.

Unité de traitement : Automate programmable, ordinateur ou régulateur.

Consommateur : Station qui a besoin de lire une variable.

Producteur : Station qui fournit une variable aux autres stations.

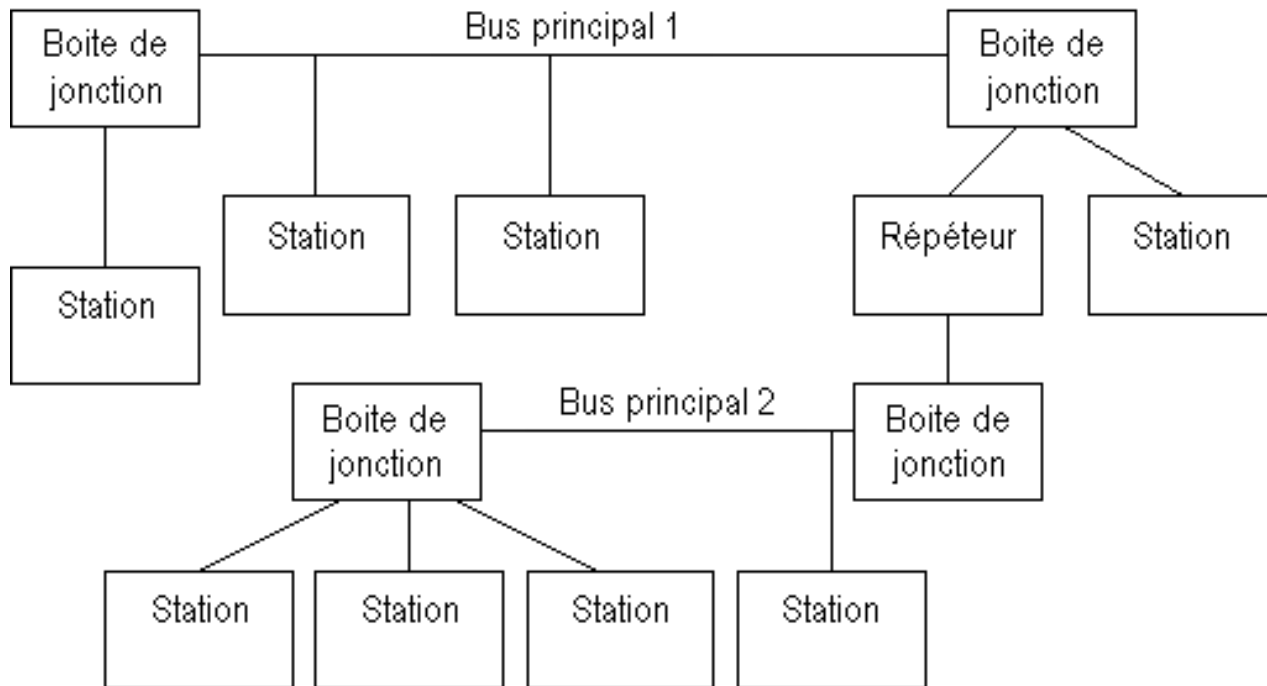
Identifieur : Adresse globale d'une variable.

Arbitre de bus : Chef d'orchestre du bus, sans son accord les stations ne peuvent ni émettre des informations ni en recevoir.

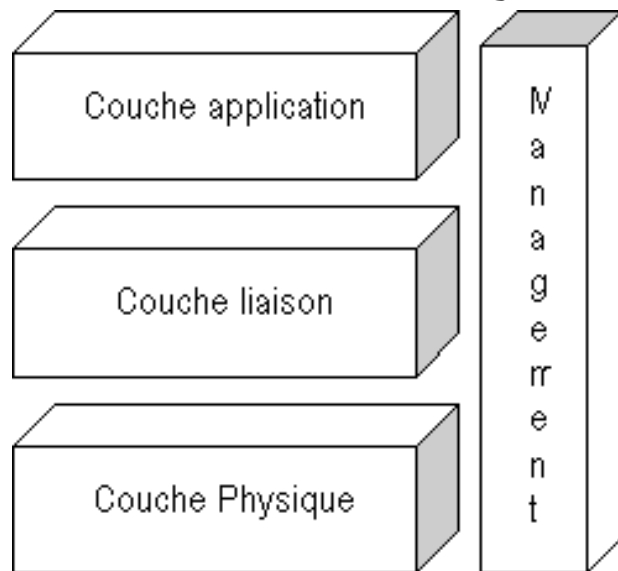
Topologie

La topologie réseau est une topologie bus classique, on peut cependant par le biais de répéteur cascader d'autre bus on obtient ainsi plusieurs bus principaux. On peut aussi y ajouter des dérivateurs multiples appeler boîte de jonctions pour connecter des éléments en un point unique sur le bus principal ou secondaire.

Exemple de topologie :



Les différentes couches du protocole Le protocole WorldFip est composé 3 couches de communications, une couche physique (physical layer), une couche liaison (data link layer) et une couche application (application layer). Il peut être complété d'une couche transverse assurant le management du réseau.

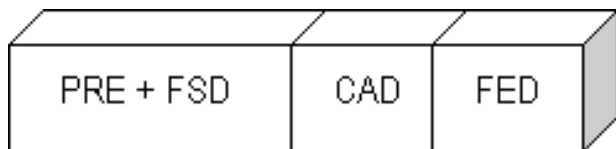


Médias utilisés

WorldFip utilise deux types de média comme support de transmission, la paire torsadé ou la fibre optique. La fibre optique à pour avantage de fournir des débits élevé, d'avoir une forte immunité aux parasites et peut s'étendre sur de grandes distances, par contre elle a pour inconvénient d'être plus fragile et plus coûteuse.

Description du protocole

Composition d'une trame



Toute trame WorldFip est constituée de trois parties :

-La séquence de début de trame (PRE + FSD)

PRE : Donne au récepteur le moyen de se synchroniser sur l'horloge de l'émetteur.

FSD : Délimite le début de trame.

-Le champ contrôle et données (CAD)

Champs contenant le contrôle et les données.

-La séquence de fin de trame (FED)

FED : Délimiteur de fin de trame.

Condition à respecter :

WorldFip peut être comparé a une base de données distribuée, le procédé utilisé pour mettre à jour les variables garantit une cohérence des données.

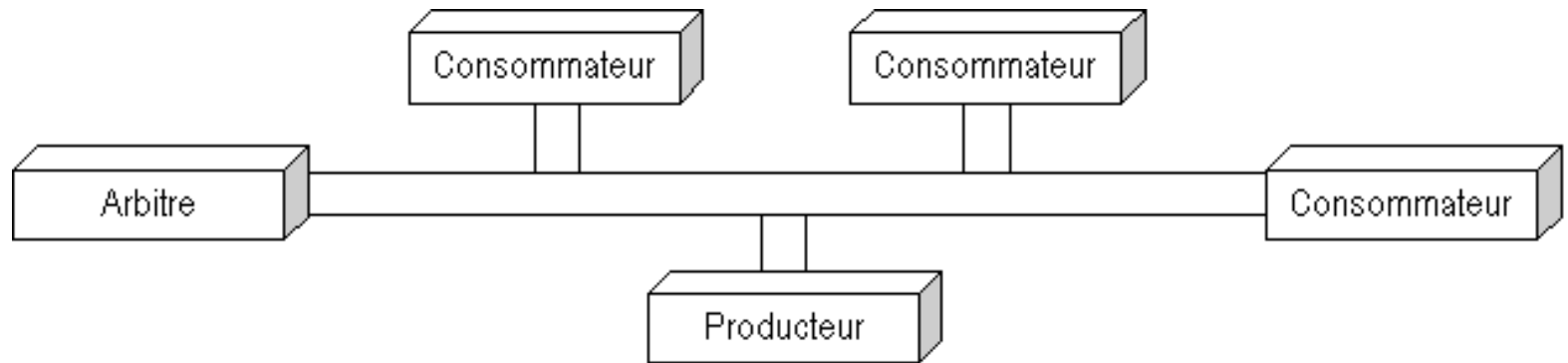
Une variable possède un identifieur unique sur tout le bus. L'identifieur des variables étant codé sur 16 bits on peut avoir jusqu'à 65536 identifieurs sur le bus.

Une variable ne peut être produite que par une seule station et ce indéfiniment dans le temps, les autres stations ne pourront y accéder qu'en lecture seule. Il ne peut y avoir qu'un seul arbitre de bus sur tout le bus à un moment donné. Le rôle d'arbitre est assuré par une station. La fonction arbitre de bus est assurée par le service ABAS de la couche application. Il existe deux méthodes d'échange de données, une cyclique, est une autre acyclique. Il existe aussi deux types d'échange, un échange de variable et un échange de message. Les messages étant plus longs que les variables. L'échange de variable cyclique et acyclique est assuré par le service MPS de la couche application. L'échange de messages est assuré par le service subMMS de la couche application. Lors de la configuration du bus, une liste des variables est établie. Pour chaque variable on définit qui en est le producteur et à quelle périodicité la variable sera utilisée sur le bus. Cette liste est détenue par l'arbitre du bus.

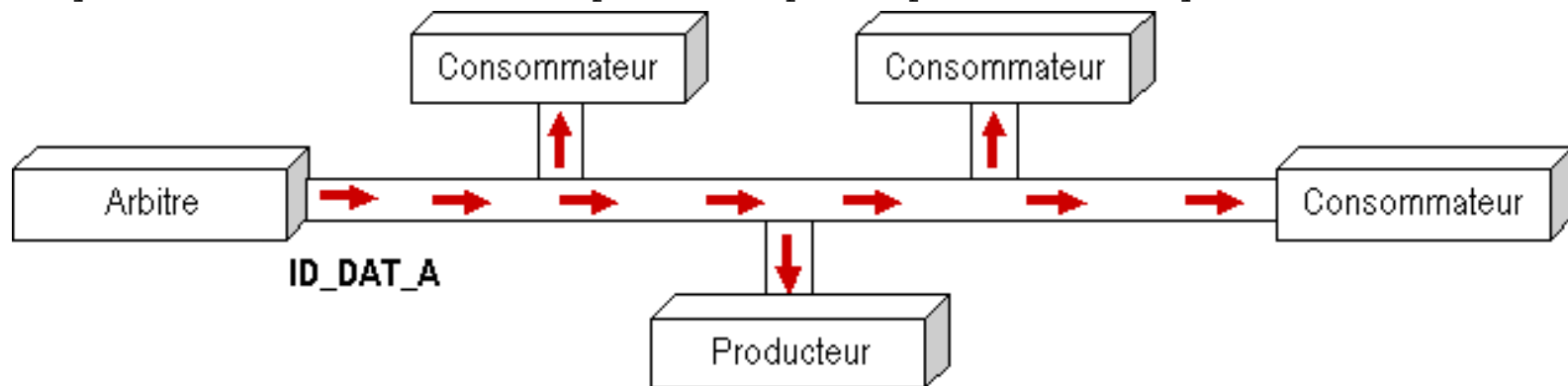
Variable	Périodicité en ms	Type	Temps en µs
A	5	OSTR_32	190
B	10	INT_8	150
C	10	INT_16	200
D	15	INT_8	160
E	20	OSTR_32	320
F	25	OSTR_32	200
G	30	INT_8	140

Cas d'un échange de variable cyclique :

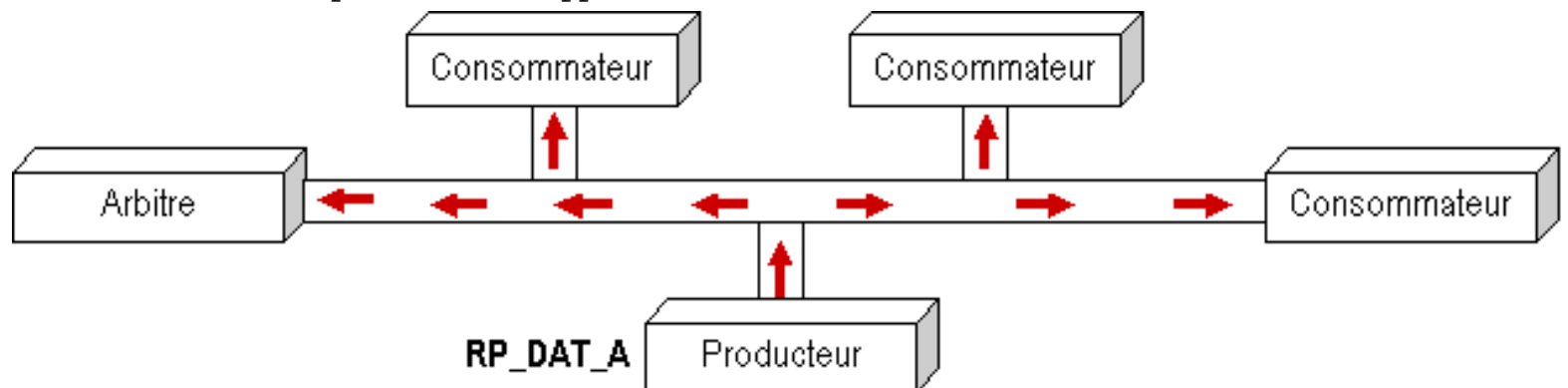
Les schémas utilisés dans les exemples comporteront un bus avec 5 stations comme ci-dessous :



L'arbitre envoie « par diffusion » une trame « Question » sur le bus pour indiquer qu'il veut lire le contenu de l'identifieur « A », toutes les stations reçoivent le message mais comme il ne peut y avoir qu'un seul producteur pour chaque variable seule la station producteur peut répondre à cette question.

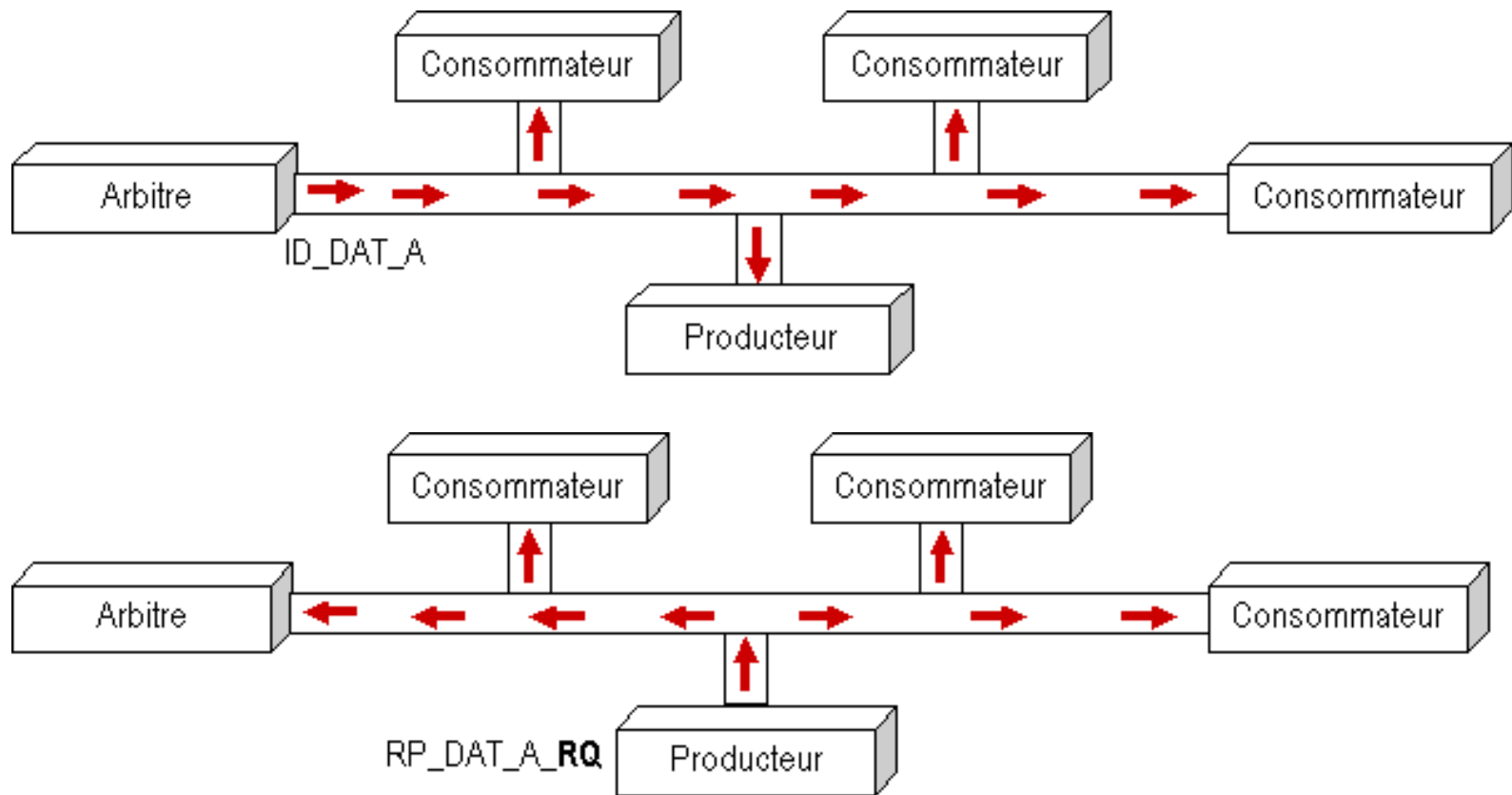


Le producteur donne sa « réponse » en envoyant la valeur de l'identifieur, toutes les stations reçoivent la valeur, les stations intéressée (consommateur) peuvent consommer cette variable ce procédé « s'appelle transfert de buffer ».

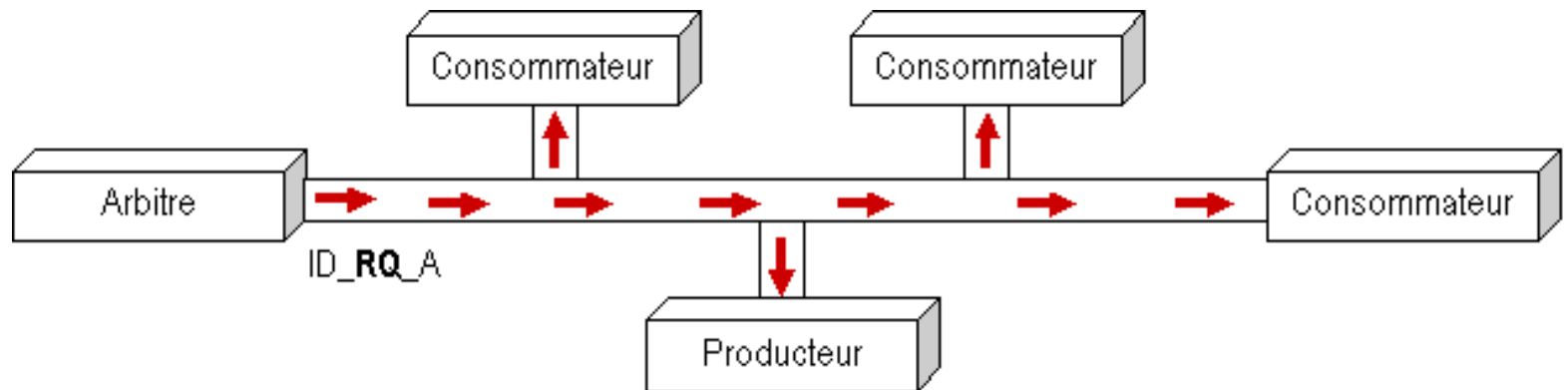
Cas d'un échange de variable acyclique :

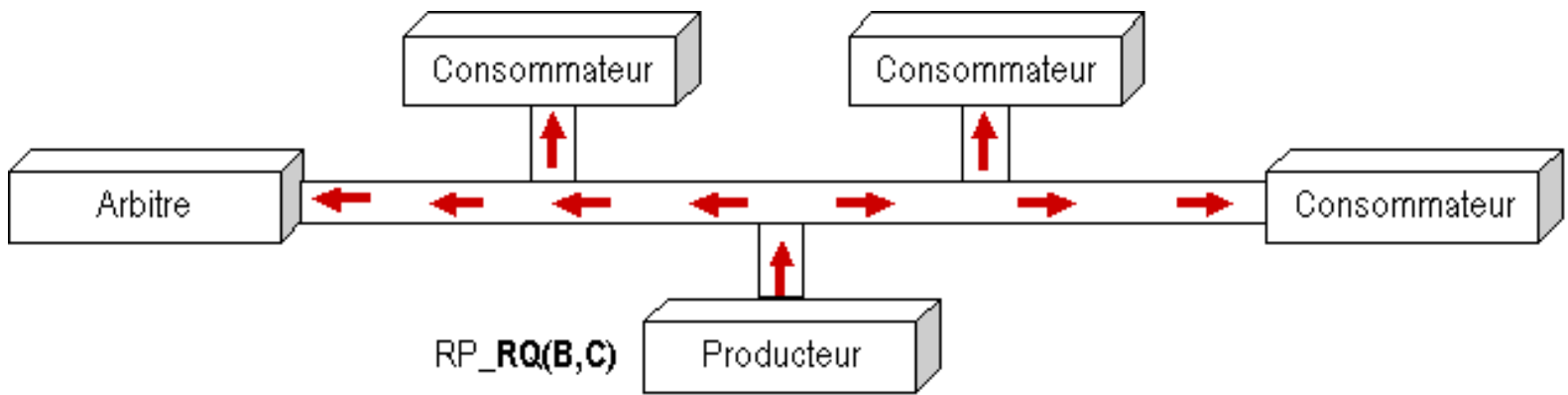
La demande de transfert de variable acyclique se fait en utilisant le procédé de transfert de variable cyclique. Comme précédemment l'arbitre envoie sa question, quant à lui, le producteur répond mais en positionnant un bit dans la trame indiquant qu'il a une

demande de requête acyclique à formuler. On peut si nécessaire ajouter un drapeau de priorité urgent.



Maintenant, l'arbitre sait que la station qui vient de répondre a une ou plusieurs variables acyclique a transférer. Il retransmet la même question que précédemment en ajoutant un bit requête pour indiquer à la station de lui transmettre la liste des variables à demander. La liste peut contenir jusqu'à 64 identifiants.

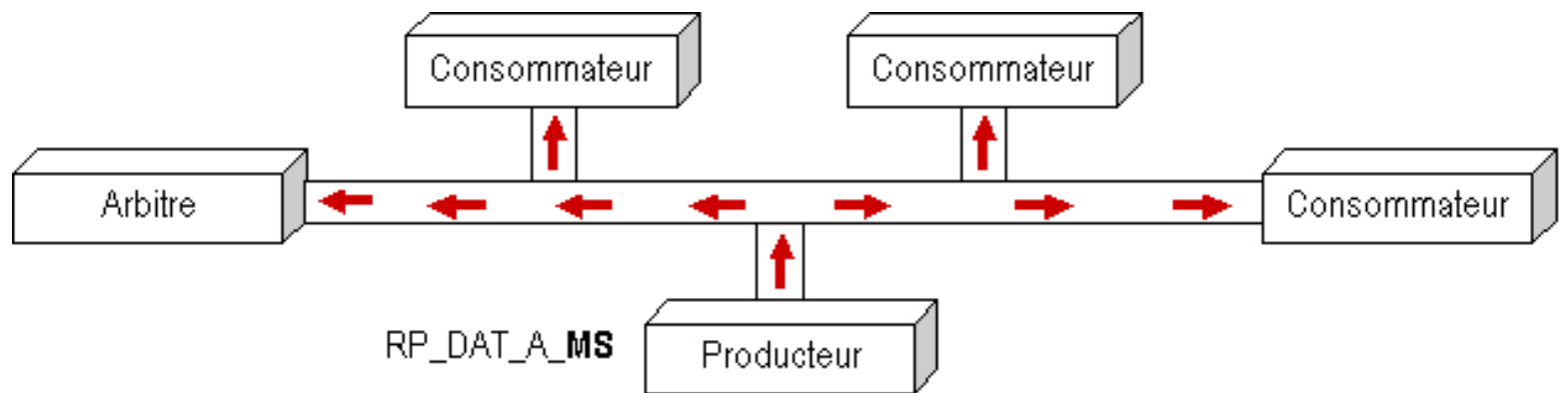




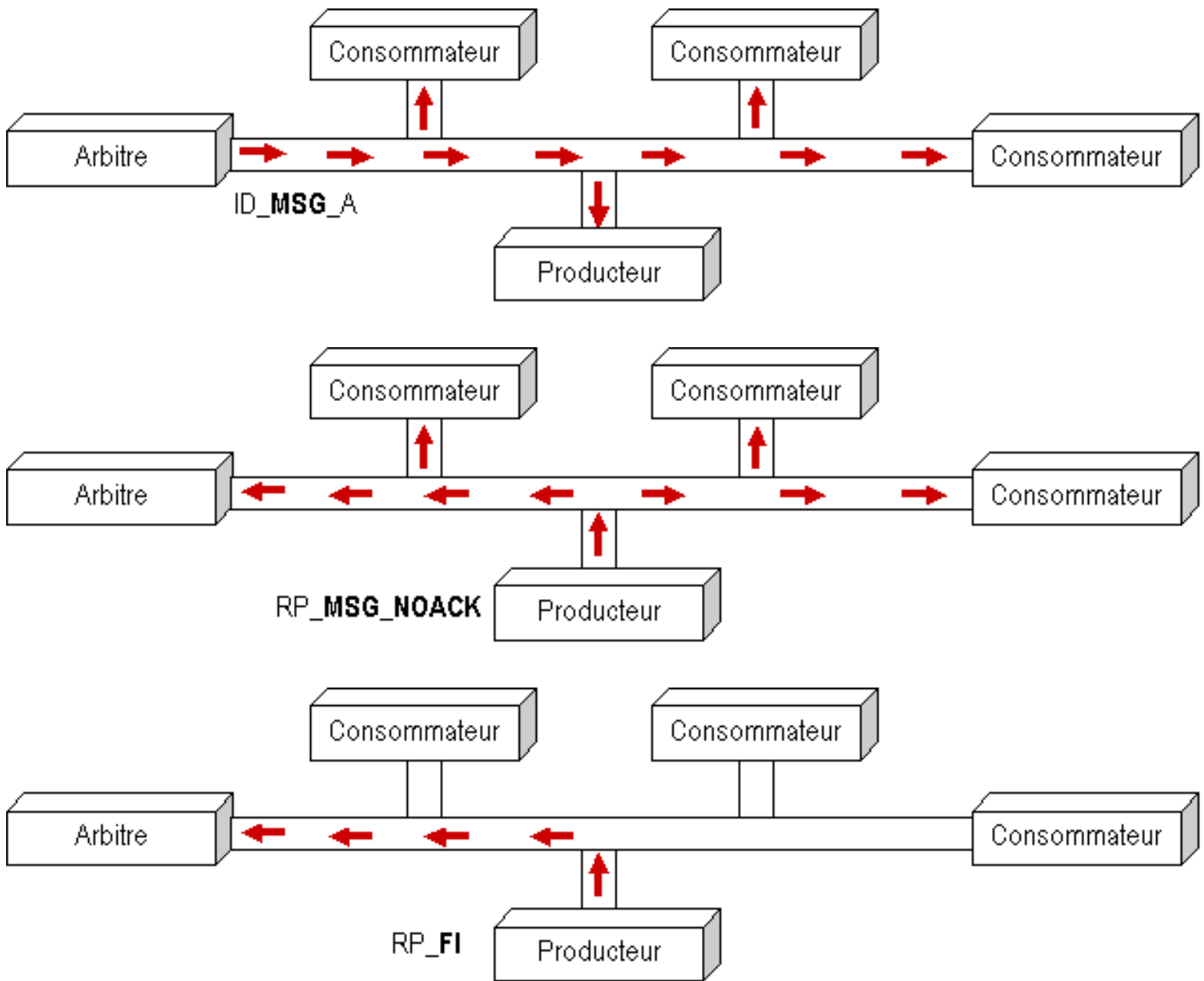
Une fois que l'arbitre reçoit la réponse, il demande la lecture d'une des variables de la liste, le processus continue ainsi comme pour une variable cyclique. Seul un producteur peut demander le transfert de variable acyclique, par contre il peut demander la lecture de variables dont il n'est pas le propriétaire. Il est possible de faire une demande spécifiée. Le producteur choisi de déclencher sa demande de transfert acyclique sur production d'une variable spécifique.

Cas d'un échange de message sans acquittement :

Les transferts de messages sans acquittement se font de façon acyclique, ils s'adressent à une station ou à plusieurs stations. Le début du processus de transfert de message se fait comme pour celui d'une variable acyclique, la différence commence quand le producteur répond, il positionne un bit indiquant qu'il a un message à transférer.



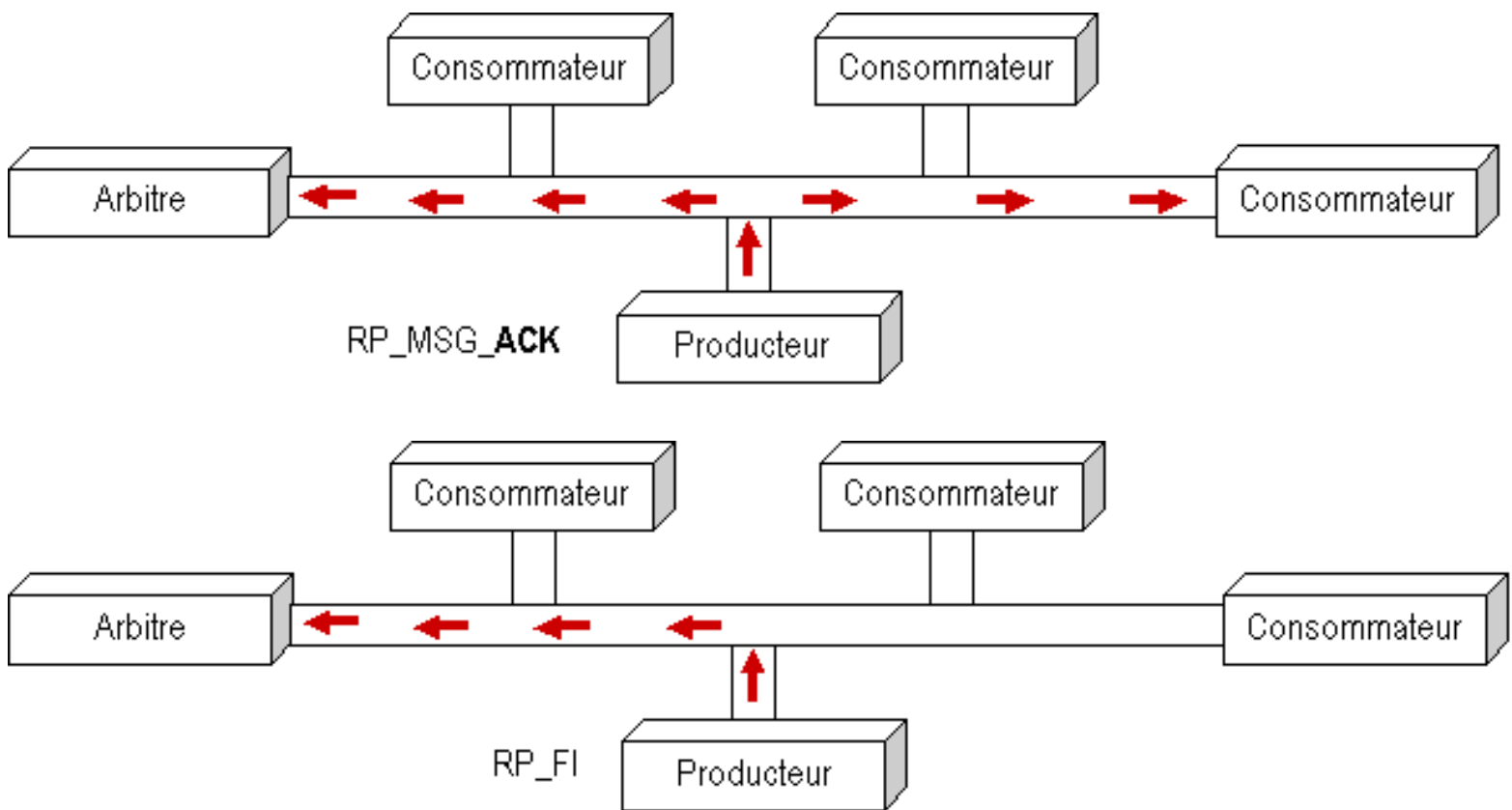
L'arbitre indique au producteur qu'il peut adresser son message et attend la trame de fin de transfert. Le producteur émet son message en précisant s'il attend ou pas un accusé de réception. Une fois celui-ci émit, il indique à l'arbitre que le transfert est terminé.



Pour éviter que le transfert s'éternise, l'arbitre arme un timer.

Cas d'un échange de message avec acquittement :

Les transferts de messages avec acquittement se font de façon acyclique, ils s'adressent à une station. Le début du processus de transfert de message se fait comme précédemment, la différence commence quand le producteur envoie son message il positionne un bit indiquant que le message est avec accusé de réception.



Un message perdu ne peut être ré-émis que deux fois. L'arbitre dispose toujours d'un timer pour éviter d'être bloqué par un transfert trop long.

Classe de services :

Tous les équipements ne peuvent pas assurer tous les services énumérés ci-dessus, pour avoir un aperçu rapide de leurs possibilités, les services ont été regroupés par classe :

Services / Classes	1	2	3	4	5	6	7	8	9
Transfert de buffer	X	X	X	X	X	X	X	X	X
Ecriture de buffer	X	X	X	X	X	X	X	X	X
Lecture de buffer		X	X	X	X	X	X	X	X
Demande explicite libre				X	X			X	X
Demande explicite spécifiée			X		X				X
Message sans acquittement						X	X	X	X
Message avec acquittement							X	X	X

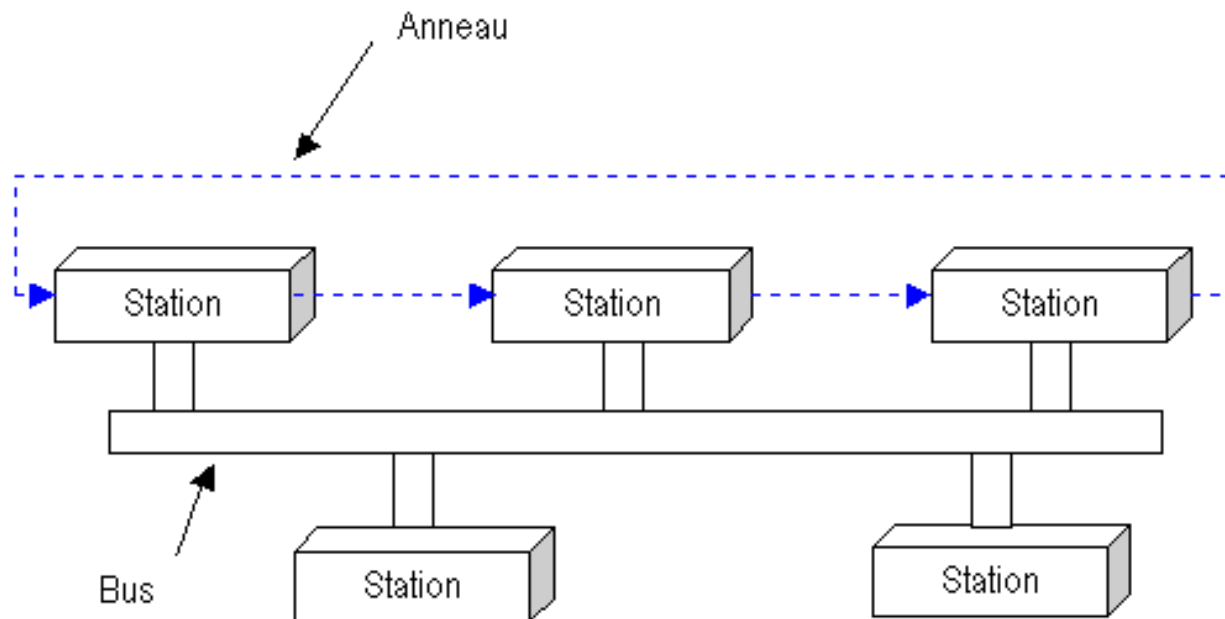
Fonctionnement du bus PROFIBUS

PROFIBUS est composé de deux protocoles de transmission appelés profils de communication DP (Decentralized Periphery) et FMS (Fieldbus Message Specification) et de profils applicatif tels que PA (Process Automation), BA

(Building Automation) ou Profisafe. Ces profils de communication définissent la façon dont les données sont transmises sur le support physique. Le profil de communication FMS est destiné à la communication entre équipements dits intelligents tels que des ordinateurs ou des automates, tandis que le profil DP est plutôt destiné à la communication entre équipements intelligents et équipements basiques tels que des capteurs et actionneurs

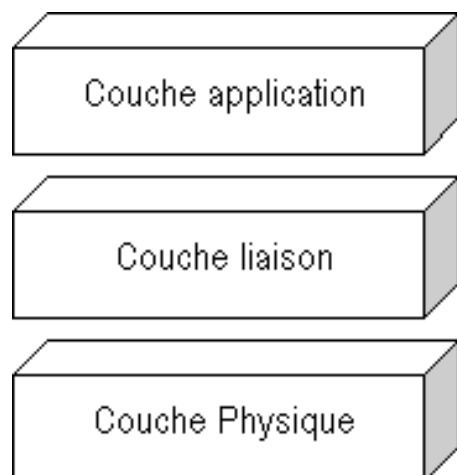
Topologie

D'un point de vue physique, la topologie utilisée est celle d'un bus mais d'un point de vue logique on a en plus une topologie anneau.



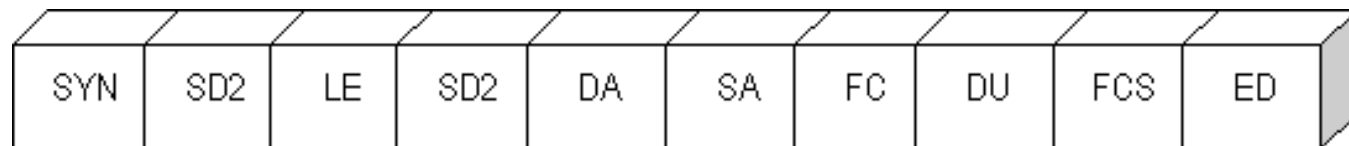
Les différentes couches du protocole

Le protocole PROFIBUS est composé de 3 couches de communications, une couche physique (physical layer), une couche liaison (data link layer) et une couche application (application layer).



Médias utilisés

PROFIBUS peut être utilisé avec trois type de média, la paire torsadée (RS485), le bifilaire (CEI 1158-2) et La fibre optique, cette dernière à pour avantage de fournir des débits élevé, d'avoir une forte immunité aux parasites et peut être utilisée pour des grandes distances, par contre elle a pour inconvénient d'être plus fragile et plus coûteuse. A noter cependant que le futur de PROFIBUS prévoit sont utilisation sur des infrastructure déjà existantes tel que le réseau Ethernet.

Description du protocoleComposition d'une trame

Syn = Synchronisation
 SD2 = Délimiteur de début
 LE= Longueur
 DA= Adresse destination
 SA= Adresse source
 FC= Code fonction
 DU= Données
 FCS = Checksum
 ED= Délimiteur de fin

Il existe deux types de stations sur le bus, les stations maîtres appelée aussi stations actives et les stations esclaves appelée également stations passives. Toutes deux peuvent recevoir ou émettre des messages, cependant, seul le maître peut émettre librement tandis que l'esclave n'émet que sur demande du maître. Il existe aussi deux méthodes de communications différentes, une réservée pour la communication entres les maîtres appelé « méthode du jeton » et une autre réservée pour la communication entres les maîtres et leurs esclaves appelé « mode maître-esclave ». Pour communiquer entres elles les stations ont la possibilité de diffuser leurs messages en mode point à point, en mode broadcast (diffusion générale) ou en mode multicast (diffusion sélective)

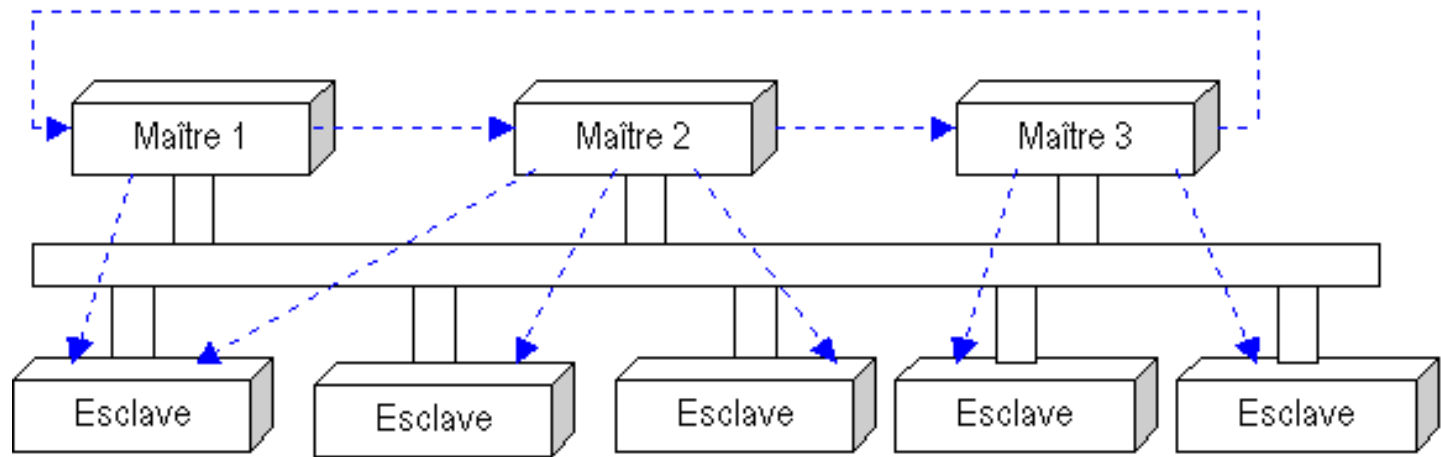
Méthode du jeton

Il ne faut pas que certain maître ait le monopole de la parole, pour cela il faut instaurer un système garantissant dans un temps défini que chaque maître puisse avoir la possibilité de parler. Les maîtres connectés au bus forment un anneau à jeton virtuel et son classées par rapport à leur adresse de la plus petite à la plus

grande. C'est au démarrage du réseau que la liste des maîtres est élaborée, ensuite le maître possédant l'adresse la plus petite acquiert le jeton en premier, il peut ainsi communiquer avec les autres maîtres ou avec ses esclaves ou tout simplement passer son tour, il retransmet ensuite le jeton au maître ayant l'adresse supérieure. La liste est mise à jour automatiquement lors d'un retrait ou d'une insertion d'un maître. Une station maître possédant le jeton est une station active.

Mode maître-esclave

C'est sur ordre de leurs maîtres que les esclaves reçoivent ou transmettent des messages.



Profil de communication DP

Le profil de communication DP destiné à la communication entre les maîtres et les esclaves est utilisé pour des échanges de données courts (244 octets) et rapide. Il peut aussi bien se faire de façon cyclique que de façon acyclique. La connexion et la déconnexion d'éléments DP s'effectuent de façon simple et rapide (Plug and Play). Il peut y avoir un ou plusieurs maîtres sur le bus. En configuration mono-maître, l'unique maître s'occupe de tous les esclaves présents sur le bus. En configuration multi-maîtres, les maîtres peuvent avoir des esclaves communs, cependant le droit en écriture sur ceux-ci n'est pas partagé, seul un d'entre eux possède un droit en écriture sur un esclave, cette station « DPM1 » est désignée lors de la configuration. De même, il existe un maître DPM2 qui est une station de configuration, de service et d'analyse pour les différents esclaves.

Transmission cyclique : Le DPM1 possède la liste de ses esclaves et une liste des variables à transférer avec leurs périodicités. DPM1 peut s'adresser directement à un esclave, à un ensemble d'esclaves ou voire tous les esclaves en même temps. Dans les deux derniers cas, il est obligé d'assurer la synchronisation des entrées et des sorties entre tous ses esclaves. Pour cela, il dispose de deux commandes « Synchro » et « Freeze »

Transmission acyclique : Pendant le temps libre entre deux transmissions cycliques, DPM1

peut envoyer des messages acycliques, il dispose pour cela cinq fonctions, MSAC1_Read pour lire les données de l'esclave, MSAC1_Write pour écrire des données vers l'esclave, MSAC1_Alarm pour transmission d'une alarme vers le maître avec attente d'accusé de réception, MSAC1_Alarm_Acknowledge pour envoi d'accusé de réception vers l'esclave et MSAC1_Status pour informer le maître de l'état de l'esclave.

Profil de communication FMS

La priorité pour ce de profil destiné aux échanges entre maîtres est la quantité de donnée échangées au détriment de la rapidité. Le profil FMS possède un dictionnaire d'objets contenant la description, la structure, le type de données ainsi que la relation entre l'adressage interne et l'adressage externe complété par sa désignation connue par les autres stations du bus. Ces objets qui sont des objets de communication peuvent être des variables, des tableaux (suite de variables de même type), des structures (suite de variables de type différents), des domaines (zones mémoire) ou des événements (message d'alarmes).

Index	Type	Adresse interne	Désignation
10	Var	4100	Température
11	Var	3600	Pression
...			

Les profils applicatifs

Ils décrivent comment utiliser les profils de communications et les supports physiques par les applications ou par les équipements. PA : Basé sur le profile de communication DP, PA utilise le support de transmission CEI 1158-2 avec dans la plupart des cas une télé-alimentation des équipements. Grâce à des interfaces décrivant le comportement de l'équipement sur le bus et une connaissance détaillée des différents équipements du marché, on peut interchanger des équipements de marques différentes sans intervention sur l'équipement. Profisafe : Basé sur le profile de communication DP, Profisafe assure la sécurité (intégrité) des transmissions (perte de trame, erreur de séquence, ...). Il définit aussi le raccordement d'équipement de sécurité tel qu'un bouton d'arrêt d'urgence, BA : Basé sur le profile de communication FMS, BA définit les règles sur tout ce qui a attrait à l'exploitation d'un bâtiment, surveillance, exploitation, traitement d'alarme... . Ces profils applicatifs ont de par leurs règles déjà définies, pour rôle de faire baisser les coûts engendrer par une étude sur l'installation d'un bus de terrain en fonction de l'application métier.

Les bus et les normes européennes

Comme pour toute nouvelle technologie, l'arrivée des bus de terrain à souffert d'un manque de normalisation. L'émergence de nouveaux produits et de nouvelles fonctionnalités ont creusé encore plus l'écart entre le début de la normalisation et la réalité technique. Ce manque qui rendait les clients frileux malgré tous les avantages que pouvait leur apporter les bus de terrain s'expliquait par la crainte de faire le choix d'une solution propriétaire qui lierait le client à son fournisseur voire même d'une

solution qui ne serait pas pérenne ou instable. Bien sur, le retour sur investissement existe bien, cependant mettre en place une solution de ce type nécessite un investissement de départ qu'il faut prendre en compte ainsi que le maintien en vie de la solution. Pour garantir le bon fonctionnement du bus de terrain, il faut sans cesse assurer la maintenance, mettre à jour le matériel existant et ajouter de nouveaux matériels. Choisir une solution normalisée garantie une solution mûrement réfléchie, stable et pérenne. Elle garantie aussi une interopérabilité entre les matériels de différents constructeurs. Les bus de terrain FIP et PROFIBUS sont tout deux entièrement normalisés sous la norme EN 50170.

Domaines d'application adaptés à chaque bus

On peut dire que les bus Worldfip et Profibus sont vraiment en concurrence, on les retrouve dans les mêmes domaines d'application telles que l'industrie alimentaire, l'industrie énergétique, l'industrie automobile, l'automatisation de bâtiment...

Comparaison des deux bus

Bus	Support de transmission	Débit Max	Nbre maximum d'unités	Longueur maximale	Taille maximale des données
WorldFip	CEI 1158-2 fibre optique	6 Mb/s	256	50 km	Pas de limite pour les messages; 128 octets pour les variables
Profibus	CEI 1158-2, RS485, fibre optique	12 Mb/s	127	24 km	244 octets

Conclusion

Le but principal des bus de terrain est atteint par les deux bus : réduire les coûts par la diminution du nombre de câbles, par une mise en place et une administration simples et par des éléments pouvant être d'origines diverses.

En matière de sécurité, les deux bus détectent les erreurs de transmission et ont une procédure de reprise.

Ils autorisent une redondance, un seul est utilisé à la fois, l'autre servant qu'en cas de défaillance.

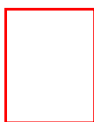
Leurs différences notables sont visibles dans les comparaisons techniques détaillées ci-dessus et essentiellement dans leurs protocoles de dialogue.

Profibus ressemble à un système de communication client-serveur entre le maître et ses esclaves où les esclaves seraient les serveurs et le maître le client. Le nombre de maîtres constituant l'anneau logique ne doit pas être trop important car il diminuerait le temps de communication disponible entre les maîtres et leurs esclaves.

Cette contrainte n'existe pas dans Worldfip.

Le recouvrement de la perte d'un maître dans Profibus n'est pas soulevé dans les documentations, on peut se poser la question sur le devenir des esclaves appartenant au maître défaillant, Worldfip ne fonctionnant pas en maître/esclave le point de défaillance pourrait être l'arbitre de bus mais la perte de celui-ci est prévue, en effet il existe sur le bus plusieurs arbitres capable de prendre le relais mais un seul peut être actif à un moment donné, le basculement se fait sur défaillance de l'actif par un algorithme d'élection embarqué qui se fait en fonction du poids de l'adresse.

Sans s'intéresser à la mécanique, d'un point de vue utilisateur, le choix de Profibus ou de Worldfip n'aura pas trop d'impact. Par contre pour une personne plus technique le choix d'un des deux bus se fera plus ressentir causé par un fonctionnement différent.



Conservatoire national des arts et métiers
Architectures des systèmes informatiques
ANNEXE 05
Comparaison des bus WorldFip et Profibus
Année 2002-2003

Conservatoire national des arts et métiers
Architectures des systèmes informatiques
ANNEXE 06
Origines et situation de quelques systèmes d'exploitation
Année 2002-2003



Ce paragraphe n'apporte rien à notre propos principal relatif aux architectures. Le lecteur qui connaît déjà ces points d'histoire, petite ou grande, ne lira pas cette annexe, celui qui ne les connaît pas pourra la lire comme un complément au chapitre 1.

Les systèmes d'exploitation ont tous une histoire et leur origine est parfois une création d'outils sans intention commerciale par un ou quelques individus insatisfaits de l'existant (Unix, Linux (1991))

Le premier vrai système d'exploitation fut celui de l'IBM 704 en 1954, dû à Gene Amdahl.

1. MULTICS, LE GRAND ANCÊTRE

La première et pratiquement seule tentative d'écriture d'un «grand système d'exploitation» à vocation universelle est connue sous le nom de *Multics* (multiplexed information and computing service).

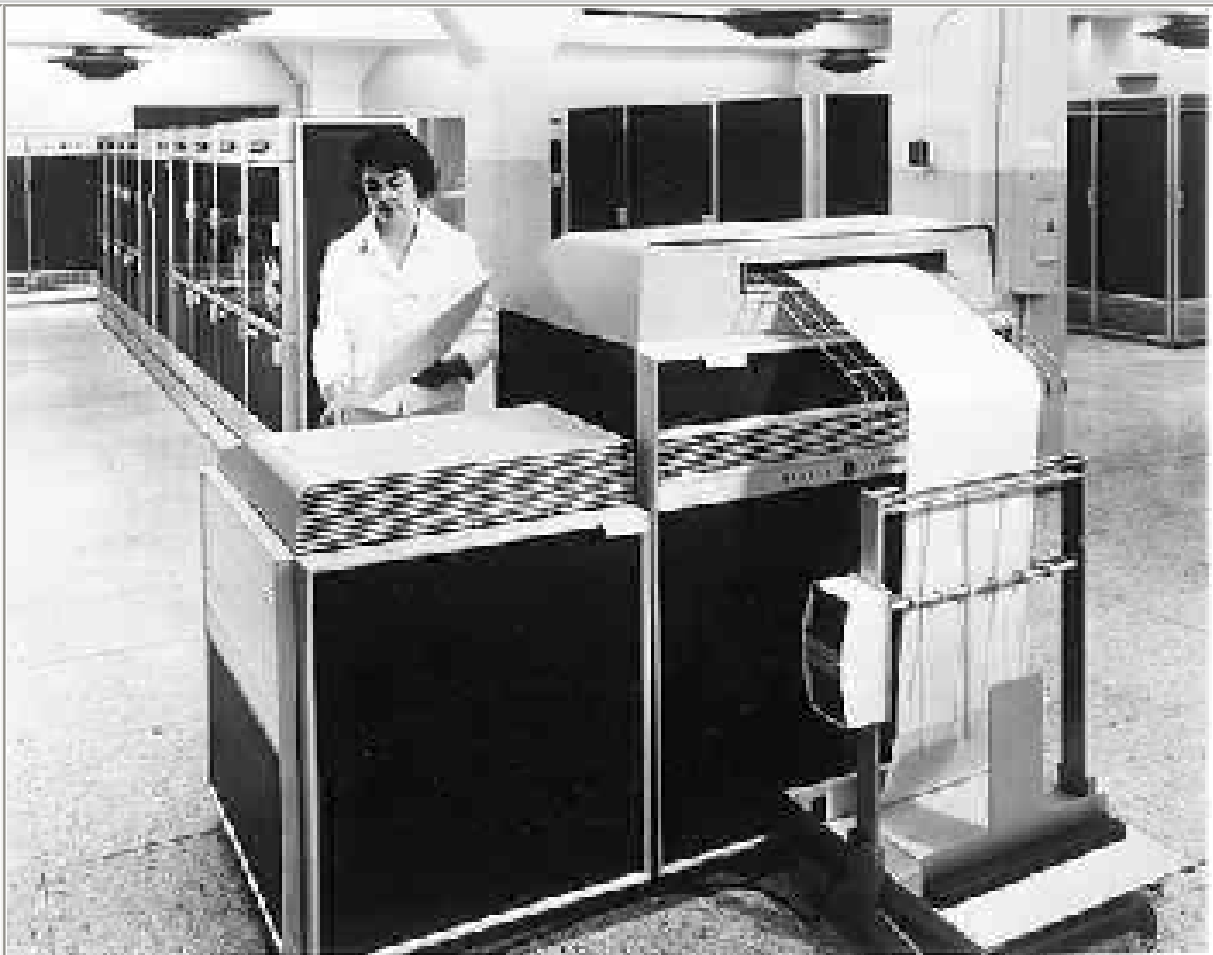
Il a été précédé par *CTSS* (compatible time sharing system) créé par Fernando Corbato au MIT utilisé de 1963 à 1973 dans cet institut. Il est installé d'abord sur une machine en fin d'existence l'IBM 709, puis sur des IBM 7090 et 7094 en 1961. La première démonstration est faite avec trois utilisateurs connectés. Chacun a l'impression d'avoir la machine pour lui seul.

L'ARPA (advanced research project agency) propose alors de financer un projet nommé MAC (machine aided cognition), de mauvaises langues ont écrit «Minsky against Corbato» bien que ce dernier ait participé au projet. Le nom Multics arrivera plus tard. Le projet est décrit dans six communications présentées au Fall Joint Computer Conference de 1965. Le but du projet est de créer un système d'exploitation pour ordinateur parfaitement fiable, capable de tourner 24H sur 24, 7 jours sur 7, d'être utilisable par de nombreuses personnes simultanément et capable en même temps d'exécuter des tâches de fond.

Les trois auteurs de Multics sont : les laboratoires Bell de 1965 à 1969, le MIT via le projet MAC, le département des ordinateurs de General electric, acquis ensuite par Honeywell information systems.

La série *GE600* est choisie comme premier support. Le *GE625* sort en avril 1965, le *GE635* en mai 1965, il est transformé en *GE645* en 1967 par ajout

- de la pagination;
- de la segmentation;
- d'un système d'entrées et sorties unique a été conçu pour gérer des dispositifs divers, disques, bandes, télétypes, sans que cela entraîne des interruptions excessives du processeur;
- d'un tampon de 4 millions de mots de 36 bits incorporant une file d'attente



câblée pour
gérer le
trafic de
pages.

Ci-contre une vue du
GE625

Les objectifs détaillés en termes de services sont :

1. l'utilisation normale se fait par accès distant;
2. le service est continu comme l'alimentation électrique et le service téléphonique;
3. les capacités peuvent être étendues ou restreintes sans réorganisation ni du système ni des utilisateurs;
4. le système de fichiers est suffisamment fiable pour que les utilisateurs n'aient besoin que d'un exemplaire du fichier qui contient leurs données et programmes;
5. les données sont partageables;
6. la fourniture des services est équitable entre petits et grands utilisateurs;
7. des environnements de programmation et des interfaces pour utilisateurs différents peuvent cohabiter;
8. le système sera «flexible» pour évoluer en fonction de la technique et des demandes des utilisateurs.

Pour cela, le stockage logique et l'administration du système ont une structure hiérarchique. La gestion multiprocesseurs a été ajoutée ultérieurement.

Note : *Multics est écrit en PL/1.* Sauf omission de notre part, c'est la première et dernière fois qu'un système d'exploitation est directement écrit dans un langage bien structuré, ce que C ne sera pas pour Unix. Une tentative ultérieure, a été la réécriture d'Unix en pascal entreprise par l'IRIA sur fonds publics, sous le nom de projet Sol au début des années 1980. Il ne s'agissait plus alors d'écrire un nouveau système.

Il s'est agi de la plus grande tentative de centralisation en informatique sous une forme explicitement analogue à la production et à la distribution de l'énergie électrique. Les concepteurs prévoyaient de câbler les villes et d'installer à la demande des prises donnant accès à l'ordinateur central fournisseur de temps de calcul. Le lecteur peut penser qu'il s'agissait d'une mauvaise idée, et pourtant, ne l'a-t-on pas reprise sous une autre forme dans le trop fameux [plan câble](#)?

Les difficultés importantes de mise au point du système d'exploitation ont conduit à ne pas donner à ce type d'idée la suite espérée par ses concepteurs, bien que de nombreuses machines aient été équipés de *Multics*, notamment des machines Bull. L'abandon de la maintenance de *Multics* par ce constructeur a d'ailleurs fait quelque bruit dans les années 1980.

Un autre système a été ambitieux, celui du projet TSS/360 d'IBM (time sharing system).

2. UNIX

Les laboratoires Bell perdent toute illusion sur le projet Multics et se retirent en 1969. Peut être par réaction, Ken Thompson, chercheur aux Laboratoires Bell, commence à écrire un système mono utilisateur. Il trouve (dit l'histoire) un PDP 7 inutilisé et commence à écrire à son idée ce qu'il nomme UNICS (Uniplexed information and computer service) en donnant pour motif la gestion de documents. Il le conçoit comme un gestionnaire le plus complet possible des entrées et sorties, mono-utilisateur et multitâche. Bell déposera la marque plus tard en l'écrivant UNIX.

En 1971, il récrit Unics dans une version multi-utilisateurs sur PDP 11/20 qui a 24 k de mémoire et un disque de 512 k, avec Ritchie qui vient de définir le langage C, successeur du langage interprété B jamais diffusé et beaucoup moins gourmand en ressources que PL/1. Le noyau Unics y occupe occupe 16 Ko, 8 Ko restant disponibles pour les utilisateurs. Les fichiers une taille maximale de 64 Ko.

Que se serait-il passé si Thompson avait écrit ce système en PL/1 au lieu de C ?
Probablement PL/1 aurait-il connu un développement considérable puisque le langage C a dû sa fortune à Unix.

Ce système pourra être porté sur une autre machine dès qu'un compilateur de C sera disponible sur elle. Le premier portage est fait en 1977 sur un Interdata 8/32. Unix est tout d'abord distribué gratuitement aux universités. Il devient un produit commercial en 1976 et plus sérieusement en 1979 avec la version 7, *Unix V7*. Soucieuses d'indépendance vis-à-vis de leurs fournisseurs et aussi pour des motifs d'enseignement des systèmes d'exploitation, les universités l'avaient adopté assez largement et y avaient apporté des modifications et adjonctions de leur crû. La version modifiée la plus connue est celle de l'université de Berkeley commercialisée par Berkeley software distribution sous le nom de BSD 4.2. en 1983. Cette souche a eu pour descendants 386BSD, FreeBSD, NetBSD, OpenBSD, SunOS, BSDI.

Pendant ce temps AT&T se rend compte que ses laboratoires ont probablement eu une bonne idée et diffuse UNIX System III en 1982, puis UNIX System V en 1983. Cet Unix est «ouvert». Ses spécifications sont disponibles mais soumises à redevances pour chaque constructeur qui l'adopte et peut le modifier comme il l'entend. Les versions «maison» quasi Unix (Unix-like) vont donc fleurir comme Xenix de Microsoft (1980) pour les I8086, Z8000 et MC68000, SunOs (1982), Unicos sur le quadriprocesseur Cray 2 (1985), AIX d'IBM (1986) et tant d'autres. Ceci n'est pas vraiment gênant pour les constructeurs nouveaux venus qui dépensent de toute façon moins en redevances que pour écrire un système. Les grands et anciens constructeurs sont agacés et refusent longtemps d'installer ce qui est perçu comme une ouverture vers d'autres mondes et une incitation à l'infidélité pour leurs clients. À cette époque, la Télémécanique adopte Unix et abandonne la conception de systèmes d'exploitation pour ses machines Solar. La société reclasse 80% de ses ingénieurs en développement de systèmes.

AT&T joue encore plus le jeu de la portabilité en publiant en 1986 SVID, System V Interface Definition, ensemble de spécifications qui doit rendre portables les logiciels entre tous les Unix par une simple recompilation des sources. Pour vérifier qu'un système est bien orthodoxe, il suffit de faire exécuter SVVS, System V Verification Suite.

Des utilisateurs, tiraillés entre *Unix System V*, *Unix BSD*, *Xenix* et d'autres, s'allient en 1984 au sein de l'IEEE pour définir une norme. Elle est nommée POSIX, publiée en 1986 et transmise à l'ISO. Les constructeurs jurent bien sûr d'adopter Posix mais font remarquer

qu'en 1987, AT&T s'est allié avec Sun Microsystems pour définir un nouveau System V, remarquablement adapté aux stations de travail de ce fabricant, synthèse de System V et de BSD qui sera nommé Solaris. Dans le même temps AT&T augmente le tarif des redevances. Cela tourne au tollé quasi général des constructeurs qui crient à la manipulation.

Ces remous conduisent IBM, DEC, HP, Apollo, Bull, Siemens et Nixdorf à créer OSF (Open Software Foundation), évidemment pour le bien des utilisateurs pour définir une nouvelle version dérivée de Posix; ce sera un quasi-Unix qui ne devra rien à AT&T. Le temps passe et on constate qu'OSF travaille sur une dérivation d'AIX qui est la version d'IBM. Ceci n'a bien sûr pas de rapport avec X/Open, créé en 1984 par Olivetti, Bull, Siemens, Nixdorf et ICL. Lassé ou exaspéré, AT&T crée Unix International, qui a les mêmes buts qu'OSF, avec Sun, Unisys, Amdahl, Fujitsu, NCR, Control Data, Prime mais aussi Olivetti et ICL. Les mêmes constructeurs se trouvent donc simultanément dans deux camps. (histoire arrêtée fin 1989), pour la suite on se reportera à l'histoire contemporaine.

Unix est toujours un système passionnant, très complet, hétéroclite et sans norme sous nombre de variantes. Le nom Unix est déposé et disponible sous licence.

3. MS DOS (MicroSoft Disk Operating System)

Le système phare des micro-ordinateurs à 8 bits des années 70 était *CP/M* de Digital Research. En notoriété et en diffusion *MS-DOS* a été son successeur pour les ordinateurs personnels, mais il n'en provient pas. Les systèmes dérivés directement de *CP/M* furent *CP/M 86* puis *Concurrent DOS*, et *DR DOS*, tous produits par Digital Research.

Sa toute première version a été écrite en 1979 sous le nom de 86-DOS par Tim Patterson alors qu'il n'existait pas de système d'exploitation pour le 8086 d'Intel sorti en 1978. Microsoft recevant la commande d'un système d'exploitation pour le futur PC d'IBM au milieu de l'année 80, achète 86-DOS ou SCP/DOS, lui ajoute une gestion de disquettes et décide de lui faire gérer 640 ko de mémoire centrale. À cette date, cela représentait dix fois la capacité normale des machines et semblait considérable. Le premier PC d'IBM, issu du projet Chess, sort en 1981 avec 64k de mémoire centrale et *PC-DOS 1.0* qui gère les deux lecteurs de disquettes prévues par IBM : 5 pouces 1/4 simple face, simple densité, 160 ko de capacité par disquette. Les dernières machines à 8 bits sous *CP/M* ont alors des disquettes de 700K et un disque de 5 ou 8 Mo. De 1981 à 1983 sortent les versions MicroSoft-DOS *MS-DOS 1.25* et *PC-DOS 1.1* qui gèrent des disquettes de 360 ko.

En 1983 sort le *PC/XT (eXTended)* avec un disque. *PC-DOS* et *MS-DOS* gèrent alors des répertoires arborescents à la manière d'Unix (ce qui exige toujours la frappe de deux ou trois touches pour le \ sur les claviers français). Apparaît également la tache de fond *PRINT*.

En 1984 *PC-DOS 2.1* et *MS-DOS 2.11* prennent en compte les usages nationaux en matière de date et de séparateurs. *DOS 3.0* gère les disquettes à haute densité de 1,2 M des *PC/AT* puis 3.1 le réseau *MS/Net*.

En 1985 lancement de *Windows 1.0*, deux ans après son annonce. C'est une bien pâle imitation du système du Macintosh. Ce ne sera pas un succès jusqu'à la sortie de la version 3.0.

En 1986 sort la version 3.2 fortement boguée puis 3.21, qui reconnaissent les lecteurs des

disquettes 3 pouces et demi de 700K des machines portables.

En 1987 les versions 3.3 gèrent les lecteurs de disquettes 3 pouces 1/2 de 1,44 Mo et doivent créer des partitions en disques logiques dans les disques de plus de 32 Mo. Mais cette limitation à 32 Mo est déjà fortement ressentie. Rien n'est fait pour dépasser la limite des 640 K de mémoire centrale, autre limitation devenue perceptible. Les astuces comme LIM (Lotus Intel Microsoft) ou EMS ne sont que des pis aller. Peut être ce sur place est-il dû à la préparation d'OS/2 considéré comme le système de l'avenir.

En 1988 MS-DOS continue seul avec la version 4.01 autant sinon plus boguée que la version 3.2, plus gourmande d'espace que les précédentes à l'intérieur des mêmes 640 K, mais gérant jusqu'à 2 Go de disque d'un seul tenant, dotée d'un interface semi graphique, et acceptant Windows en 1990 ce qui rend OS/2 encore moins attrayant.

En 1991 sort la version 5 et en 1994 la première version 6.

4. CHEZ APPLE

En 1977, système d'exploitation du Apple II est Pro-DOS, rudimentaire et efficace.

En 1983 apparaît le système d'exploitation de la Lisa et du Macintosh. Il était plus qu'inspiré de l'Alto et du Dorado visibles à Xerox PARC que Steve Jobs avait longuement visité. Il avait un interface à icônes avec fenêtres et menus déroulants actionnés par une souris. Les systèmes de Xerox étaient elles-mêmes issues de travaux des années 1960 menés par le professeur Engelbart à l'université de Stanford qui a inventé la souris, l'interface graphique, la conférence assistée par ordinateur, le traitement de texte, et a mis au point le premier hypertexte en 1963, etc. Ce système sera porté l'année suivante sur le nouveau Macintosh. Il est toujours considéré comme l'exemple auquel les autres sont comparés.

5. OS/2 Operating System number 2

Il est issu d'une coopération entre IBM et Microsoft. Il a été annoncé en 1986 et livré à partir de 1987, séparément par Microsoft et IBM. Il apportait deux réponses aux besoins exprimés par les utilisateurs de MSDOS, la première était la capacité de gérer plus de 640 Ko de mémoire centrale, la seconde était d'être multitâche bien que mono utilisateur. En addition il fournissait, ce n'était pas une exclusivité, un interface graphique. Les décideurs ne sont pas vraiment convaincus de sa suprématie et retardent leur décision de passer de MSDOS à OS/2 ou à Unix en attendant mieux.

6. CPM 86, CONCURRENT DOS et DR DOS

C'étaient les vrais successeurs de CP/M (Control program monitor) écrit pour le Z80 par Gary Kildall en 1976; et de MP/M multiutilisateur en 1978, tous les deux produits par Digital Reaseach. Ils fonctionnaient tous les deux sur des microprocesseurs à 8 bits.

Quelles que fussent leurs qualités, leur succès commercial a été très faible. Cela probablement pour deux raisons : la première est l'incompatibilité des processeurs de la gamme 80xxx d'Intel avec les précédents 8080 et 8085. Les utilisateurs n'avaient alors

aucune raison de rester attachés au même fournisseur de systèmes, la seconde est le fait qu'IBM ne les a pas choisis.

La petite histoire dit qu'IBM convoqua un 24 décembre les responsables de Digital Research qui, soit par impossibilité, soit par humeur reportèrent le rendez vous aux derniers jours de l'année. Entre temps IBM avait passé contrat avec Microsoft pour MSDOS. Selon des sources plus crédibles, la perte du marché d'IBM est due au président de Digital Research, Gary Kildall qui n'a pas donné suite aux demandes d'IBM dans le milieu de l'année 1980, ce dont l'intéressé s'est toujours défendu, disant qu'il n'avait pas été sollicité. Une dernière version est que Gary Kendall était bloqué dans un avion en panne ou en retard et que sa femme sollicitée par IBM refusa de signer la licence en son nom.

7. PROLOGUE

Ce système fut écrit pour les *MICRAL* de la société REE, à processeurs 8008 puis 8080 et 8085, et enfin 80xxx, modifié et étendu jusqu'à être multi-utilisateurs. Cette société a été ensuite rachetée par BULL, alors que l'entretien et le développement de Prologue revenait à Prologue SA. Sa part de marché est restée faible.

8. MOS Mercure Operating System

C'était un système d'exploitation de la société Mercure, qui a un temps fabriqué des machines. Il était lui aussi multi-utilisateurs. Sa part de marché est restée faible.

9. VAX VMS Virtual Memory System

Il provient d'une extension à la gestion de mémoire virtuelle de *RSX11M*, système d'exploitation de DEC pour la série PDP11. Il a été ensuite réécrit en assembleur Vax. Contrairement aux précédents, il s'agit d'un système appartenant au constructeur.

10. LES WINDOWS

Windows a été annoncé par Microsoft en novembre 1983 en réponse au système du MacIntosh diffusé au début de la même année. La première version 1.0 n'est sortie qu'en 1985 et a été suivie très vite par plusieurs mises à jour.

La version 2.0 est de novembre 1987. Elle apportait des améliorations dans la présentation, les menus, les boîtes de dialogue. Jusqu'alors, Windows ne demandait que l'Intel 8086 ou 8088 en mode réel, avec 1 Mégaoctets de mémoire. La version Windows 386 (qui a suivi peu après) utilise le mode virtuel 86 pour gérer le multitâche.

La version 3.0 est de mai 1990. Sa nouveauté était de fonctionner dans le mode protégé des processeurs d'Intel qui existait depuis le 80286, donnant accès à 16 Mo de mémoire.

La version 3.1 est d'avril 1992. Elle était la première à gérer les polices True Type, le son, la technique OLE et les boîtes de dialogue.

Windows NT apparaît en juillet 1993. Le modèle de programmation à 32 bits d'Intel apparut

dans le 80386 est géré. Ce système est porté sur des machines Risc.

Windows 95, nommé d'abord Chicago ou parfois Windows 4.0, apparaît en août 1995. Dès le mois de janvier, il pouvait être acheté pour 40 £ en version bêta à condition que l'utilisateur signale les erreurs constatées. C'était la première fois qu'un éditeur faisait ainsi travailler ses clients. Comme NT, il gère le modèle 32-bit. Bien que n'ayant pas la sécurité (relative) et la portabilité de NT, il a l'avantage de ne demander que 4 Mo de mémoire. La compatibilité des applications n'est pas complète avec les versions précédentes.

11. LA SITUATION EN MARS 1991

La question permanente de savoir quel système dominera le marché pour les micro ordinateurs n'a toujours pas de réponse. Les «petits» : *Prologue*, *DRDos*, *Concurrent Dos* ou *MOS* n'ont plus aucune chance d'y arriver. Restent MSDOS, Unix et OS/2. La question sous une forme plus immédiate est: doit-on passer de MSDOS à l'un des deux autres. Elle est posée en raison des limitations évoquées plus haut : gestion de 640 K octets de mémoire centrale. Ceci n'a été ressenti comme une limitation qu'à partir de 1985. Le saut quantitatif à partir de CP/M avait tout de même été d'un facteur 10 de 64 Ko à 640 ko. Il est mono utilisateur et mono tâche alors que les processeurs 80486 en voie d'annonce sont crédités de 80 Mips! et que l'intégration des circuits se poursuit.

Lorsque OS/2 a été annoncé, il était bien le système de l'avenir : extension de l'espace de mémoire, multitâche, interface graphique le rendant concurrent du MacIntosh et compatible avec MSDOS pour les applications. Le prix était élevé, de l'ordre de 6000 F. au lieu de 1000 F. pour MSDOS mais il était admis que les utilisateurs franchiraient le pas et que le prix baisserait. Les faits ont joué contre cette perspective. Le prix des mémoires a crû considérablement à ce moment-là, rendant d'autant plus onéreux le passage aux 2 ou 4 Mo nécessaires. La boîte de compatibilité avec MSDOS s'avéra bien peu efficace et la compatibilité annoncée illusoire. Une revue annonçait alors sans ironie que la compatibilité était garantie pour les programmes qui ne faisaient appel ni aux entrées sorties ni au système! L'attente des clients qui souhaitaient peut-être simplement de nouvelles versions de logiciels connus fut perçue par les développeurs qui s'interrogèrent alors sur la rentabilité d'investissements d'adaptation ou de développements nouveaux.

Dans le même temps IBM et Microsoft mettaient pratiquement deux ans pour fournir Presentation Manager, qui devait faire des machines des quasi MacIntosh. Un des arguments pour le passage à OS/2 avait manqué son but. Les premières versions s'avérèrent relativement boguées, gourmandes en mémoire et lentes. Ceci incita nombre de fabricants de machines à déclarer OS/2 adapté au 80386, car le 80286 n'était pas assez puissant pour lui. Cela a certainement plus nui à OS/2 que cela n'a fait vendre des 80386.

Toujours dans le même temps Microsoft dont les recettes provenaient pour l'essentiel de MSDOS a créé Windows, qui a rendu, bien qu'un peu artificiellement, MSDOS multitâche et quasi MacIntosh. Il construisait ainsi le concurrent direct de son produit OS/2, «presentation manager» inclus.

Pour couronner le tout, courant janvier 1991, le Wall Street Journal annonçait que Microsoft se retirait de l'opération OS/2 et laissait IBM seul. Cette information était démentie vigoureusement dès le lendemain, mais le mal était fait. Au milieu de la même année, Microsoft commence à livrer MSDOS 5.0 attendu depuis longtemps mais qui contrairement à l'attente n'est pas multitâche et donc encore moins multiutilisateur.

12. LA SITUATION EN 1992

Microsoft joue maintenant la seule voie de Windows, couche placée au dessus de MSDOS, multitâche et dotée d'environnement graphique «à la MacIntosh». La dernière annonce est Windows NT (new technology) pour 1993 qui sera un système d'exploitation à lui seul et n'aura donc plus besoin de MSDOS pour être exécuté. Le succès de Windows est considérable en nombre d'exemplaires livrés. Il est devenu un vrai multitâche, il est pourvu de nombreux progiciels.

Faut-il passer à Unix ?

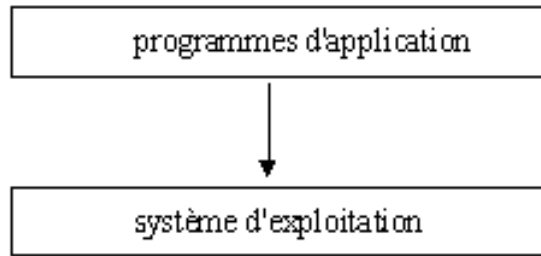
À la différence de Windows et d'OS/2, Unix est utilisé depuis vingt ans. Il est le préféré des professionnels, scientifiques et techniciens. Il est premier sur les stations de travail pour cela. Il est multitâche ET multiutilisateurs. Sa gestion de mémoire est bonne. Les applications sont aisément portables, sous réserve de recompilation bien sûr, ce qui est à la portée des développeurs et des techniciens mais pas des utilisateurs. Toutefois, les MAIS sont tout aussi nombreux. Si l'on conserve Unix tel quel, on dispose du jeu de commandes, quelque trois cents, très abrégées. Si l'on utilise, et depuis peu un interface enfin convivial, on a le choix entre OSF/Motif, Open Look, NeXT, Workspace Manager, mais cela détruit trop souvent la portabilité, y compris celles des sources.

Des initiatives existent néanmoins pour augmenter cette portabilité "pour l'utilisateur", portabilité des codes objet. Motorola a créé OPEN88 pour assurer la portabilité du code binaire entre machines basées sur le même 88000 Risc de la société. L'OSF déjà nommée propose le format ANDF Architecture Neutral Distribution Format. Les programmes conformes à cette norme s'installeraient d'eux-mêmes sur toute machine Unix.

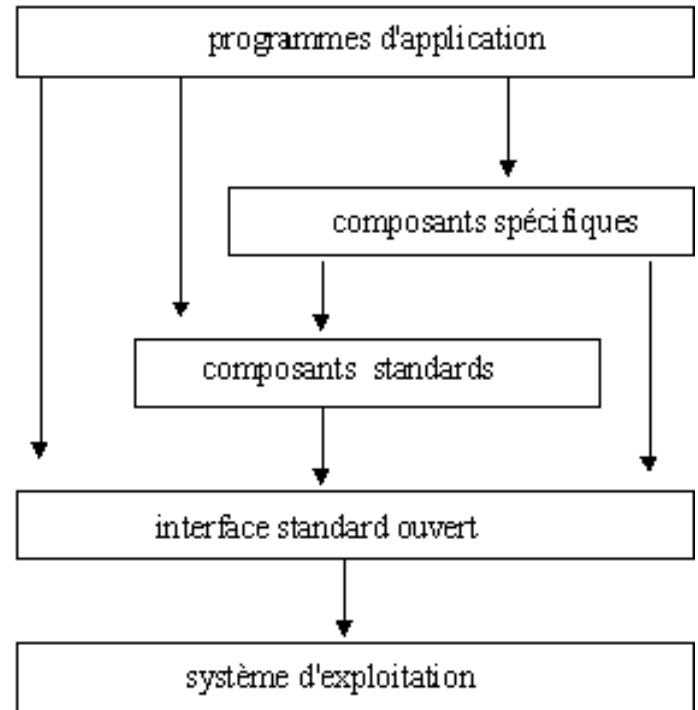
L'IEEE développe POSIX (Portable Operating System Interface X), jeu de normes d'interface.

Cette évolution de la notion d'interfaces indépendantes du système d'exploitation proprement dit qui devient essentiellement tourné vers la machine et non vers l'utilisateur peut être représentée comme suit :

SYSTEME TRADITIONNEL



SYSTEMES FUTURS



Un succès réel de cette initiative demandera au moins plusieurs années.

13. LA SITUATION EN 1999 et CONCLUSION

Le lecteur appréciera lui-même l'évolution des choses dans les cinq dernières années.

- Microsoft a gagné un quasi monopole par ses versions successives de Windows : 3.1 puis 95, NT, 98, 2000, etc.. Il soutient Intel et sa gamme x86 et est soutenu par lui. Une modification structurale importante entre Windows 3.11 et Windows95 a consisté à abandonner MSDOS comme fond du système d'exploitation. Windows95, ses successeurs et Windows NT sont ce que l'on nomme des systèmes natifs. Une nouveauté intéressante dans Windows NT consiste à utiliser la technique bien connue en matière de compilation de langage intermédiaire. Windows NT est disponible sous la forme d'une émulation du matériel nommée HAL pour «hardware abstraction layer». On peut dire que Windows NT fonctionne sur une machine virtuelle. Pour fonctionner sur un processeur quelconque Intel, Alpha, Power PC, Mips etc.. il suffit au fournisseur de traduire cette couche pour sa plate-forme.
- L'arrivée de Linux (1991), né lui aussi de l'insatisfaction d'un individu, Linus Thorvald, en concurrent des unix et déduit au départ de Minix (1985).

Le lecteur pourra utilement se demander, d'un point de vue économique, si l'informatique est semblable aux autres secteurs d'activité et comparera la durée de vie des systèmes d'exploitation, même sous des versions successives, de l'ordre de plusieurs dizaines d'années à celle des processeurs, une à quelques années.

En 2003 : IBM annonce la suppression totale du support de son système d'exploitation OS2 Warp. Il était sorti au début des années 1990 pour être rapide, stable, un vrai concurrent pour Microsoft. Malheureusement, la version 3.0 de 1994 posait un maximum de problèmes de

compatibilités logicielles et matérielles. Les versions suivantes n'avaient jamais repris les parts de marché perdues avec cette version 3. Au final, seules quelques grandes entreprises l'utilisaient encore, surtout comme applications clients-serveurs. La dernière version sortie datait de plusieurs années.

Quelques documents :

Sur Multics :

un très bel article de synthèse de Fernando Corbato et J. H. Saltzer du Massachusetts Institute of Technology, et de C. T. Clingen d'Honeywell Information Systems Inc. Article présenté au Spring Joint Computer Conference de 1972.

Sur le modèle POSIX :

«IEEE's posix : making progress», IEEE Spectrum, décembre 1991.

Sur les aventures d'Unix :

un site très complet :

<http://www.levenez.com/unix/>

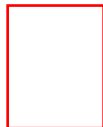
et sur une comparaison avec son concurrent Windows NT :

T. Yager, Ben Smith, «Unix est-il mort», Micro systèmes octobre 1992, pages 85 à 91.



Conservatoire national des arts et métiers
Architectures des systèmes informatiques
ANNEXE 06
Origines et situation de quelques systèmes d'exploitation
Année 2002-2003

Conservatoire national des arts et métiers
Architectures des systèmes informatiques
ANNEXE 07
Réflexions éparses et bêtisier
Année 2002-2003



1. RÉFLEXIONS ÉPARSES

Sir William Hamilton, 1866.

«Les mathématiques ne sont en aucune manière une voie conduisant à la logique, que ce soit à la logique spéculative ou à la logique pratique... Le philosophe doit avoir quelque connaissance de leur objet et de leur méthode, mais il ne doit s'y adonner qu'avec modération et précaution : un mathématicien en matière contingente est comme un oiseau de nuit en plein jour.»

Extrait de «Discussions on philosophy and litterature education and university reform»,

cité par L.G. Vidiani dans Quadrature n° 8, Magazine de mathématiques pures et appliquées, mai-juin 1991.

Lavoisier : *«L'impossibilité d'isoler la nomenclature de la science et la science de la nomenclature, tient à ce que toute science physique est nécessairement fondée sur trois choses : la série des faits qui constituent la science, les idées qui les rappellent, les mots qui les expriment. Comme ce sont les mots qui conservent les idées, et qui les transmettent, il en résulte qu'on ne peut perfectionner les langues sans perfectionner la science, ni la science sans le langage.»*

Cité par Jean Richert, Pour la Science, page 110, octobre 1994.

Confucius : *«Rectifiez les mots.»*

et *«Si l'État est en péril, refaites le dictionnaire.»*

Alonzo Church, né en 1903, enseigna à Princeton de 1927 à 1968. Il fut alors accueilli comme professeur de philosophie et de mathématiques par l'Université de Californie à Los Angeles, jusqu'à 1990. Heureuse contrée où un homme chargé d'âge, n'est pas montré du doigt au prétexte ridicule qu'il volerait le travail d'un autre en poursuivant une activité professionnelle.

Le lecteur ajoutera les contre exemples aisés à trouver : Luc Montagnier et Jean-Loup Chrétien, frappés peut-on dire, par les règlements sur la limite d'âge, suivent cet exemple aux Etats-Unis, l'un à l'université de New-York, l'autre à la

Nasa. Selon l'expression du premier : *«On ne regarde pas votre date de naissance pour faire de la recherche.»*

Gregory Chaitin, à propos de la démonstration de Kurt Gödel sur l'incomplétude des systèmes axiomatiques :

«Lorsque, jeune garçon, j'essayais de comprendre cette démonstration, je dois avouer que je pouvais lire celle-ci et la suivre pas à pas, mais d'une manière ou d'une autre, je n'ai jamais réellement éprouvé que je la saisisais véritablement.»

Bulletin of the European Association for Theoretical Computer Science (EATCS), No. 50 (Juin 1993), pp. 314-328.

Michel Volle, <http://www.volle.com> :

«Connaissez-vous la recette pour faire sérieux ? la voici :

- affecter de croire que le lecteur connaît les définitions : cela dispense de les indiquer ;*
- donner les démonstrations selon le formalisme le plus strict : il faudra du mérite pour les comprendre ;*
- se référer au cadre théorique le plus abstrait : tout ce qui est pratique est vulgaire ;*
- ne pas donner d'indication sur les intentions qui président au raisonnement : ce serait pécher contre la rigueur ;*
- sauter deux lignes sur trois dans les calculs : le lecteur ébahi cherchera le passage d'une ligne à l'autre ;*
- ne pas relire les épreuves d'imprimerie : on est au dessus de ça, et une erreur par ci par là contribue à la difficulté.»*

À propos d'abstrait : *«Combien de gens ne sont abstraits que pour paraître profonds»*

Joseph Joubert, 1754-1824, moraliste français, - Pensées

2. BÊTISIER

Commençons par le plus grand. Cicéron rapporte dans les Tusculanes, III, 28, 69, que

Aristote *«affirme que la philosophie sera bientôt tout à fait achevée»*, fragment 53, Rose.

«Je souhaiterais juste dire une boutade, mais qui est très instructive quant aux limitations de la pensée humaine. J'étais tout à l'heure dans une salle où il y avait une statue d'un grand physicien de l'époque des Lumières, qui était allé faire des mesures concernant le mètre en Bretagne, et qui en était revenu avec ses mesures. On lui avait alors posé la question suivante : "Avez-vous eu des

difficultés avec les habitants ?"

Il a répondu : "Non, pas trop, mais quel retard dans leur pensée : ils vont jusqu'à imaginer qu'il y a une influence de la Lune sur les marées !" ».

Pierre Perrier dans un débat à l'Académie des sciences morales et politiques.

«Qu'est-ce qui peut être plus absurde, manifestement, que la perspective suggérée de locomotive se déplaçant deux fois plus vite que les diligences ?»

Une revue anglaise en 1825.

«Voyager en train à grande vitesse n'est pas possible car les passagers, incapables de respirer, mourraient d'asphyxie.»

Un professeur d'anatomie en 1841.

«Le monde est aujourd'hui sans mystère».

Marcellin Berthelot, XIX^{ème} siècle, chimiste illustre. Cette phrase est la

première de son ouvrage sur «Les origines de l'alchimie» écrit en 1885.

Il était professeur au Collège de France (1865), membre de l'Académie de médecine (1863), de l'Académie des sciences (1873), et de l'Académie française (1901), 5 fois président de la Société chimique de France, sénateur à vie en 1881, ministre de l'Instruction publique (1886-1887) et des Affaires étrangères (1895-1896), Inspecteur général de l'Instruction publique (1876). Sa philosophie était positiviste, il était violemment opposé à la théorie atomique dont il bloqua l'enseignement jusqu'en 1902. Alfred Werner (prix Nobel en 1913) qui fait son stage post-doctoral dans le laboratoire de Berthelot, commenta ultérieurement l'influence de ce dernier : *«Un tel sectarisme fut plus désastreux pour la France que le traité de Francfort».*

«Le champ des inventions est épuisé».

Charles H. Duell, délégué aux brevets américains, 1899.

«Lorsque l'Exposition universelle de Paris fermera ses portes, la lumière électrique s'éteindra avec elle et l'on n'en entendra plus jamais parler.»

Erasmus Wilson, professeur à l'université d'Oxford, 1878.

En 1880, alors que l'étude de la logique s'éveillait d'un très long sommeil,

Brochard écrivait :

«La logique est une science faite. L'ère des découvertes, on peut l'affirmer sans crainte, est close pour elle.»

Dans La logique de J.S. Mill, Revue philosophique, 1880, reproduit dans Etudes de philosophie ancienne et de philosophie moderne, Alcan, Paris, 1912.

«Le téléphone est une invention étonnante, mais qui voudra un jour se servir d'un tel appareil ?»

Rutherford Birchard Hayes, président des USA de 1877 à 1881.

«Mon invention sera exploitée pendant un certain temps comme une curiosité scientifique, mais à part cela elle n'a aucune valeur commerciale quelle qu'elle soit.»

Auguste Lumière, co-inventeur du cinématographe en 1895.

«La radio n'a pas d'avenir.»

Lord Kelvin, 1897.

«La science physique est pratiquement terminée, il y a encore des constantes à mesurer, un certain nombre de travaux à faire, mais enfin toutes les grandes idées sont obtenues, il ne reste que deux petites choses à examiner et c'est l'affaire des prochaines années : le rayonnement du corps noir et l'expérience de Michelson.»

Lord Rayleigh, 1898, cité par François Le Lyonnais, Science et synthèse, collection Idées, NRF, 1967.

«Les films parlants sont bien intéressants mais je ne crois pas qu'ils restent à la mode. La synchronisation parfaite du son et de l'image est absolument impossible.»

Louis Lumière, co-inventeur du cinématographe, 1920.

«Qui diable voudrait entendre les acteurs parler ?»

Harry Warner, fondateur de Warner Brothers, 1927.

«Le cheval restera, alors que la voiture n'est qu'une nouveauté sans lendemain, une lubie.»

Directeur d'une banque conseillant à Henri Ford de ne pas investir dans la Ford Motor Company, 1903.

Réaction des militaires français qui, en 1910, avait vu fonctionner la première mitrailleuse française (Hotchkiss). Voyant que celle-ci avait complètement vidé une caisse de balles en une minute il s'écrièrent :

«Cette arme est totalement inutilisable : elle consomme beaucoup trop de munitions !»

«Les avions sont des jouets intéressants mais sans aucune valeur militaire.»

Général Foch, circa 1912.

«Il n'y a pas la plus petite indication que l'on puisse obtenir un jour de l'énergie nucléaire. Cela impliquerait de pouvoir désintégrer l'atome à volonté.»

Albert Einstein, 1932.

«Mettre un homme dans une fusée à plusieurs étages et le propulser dans le champ

gravitationnel de la Lune, d'où il pourrait faire des observations scientifiques, peut-être s'y poser sain et sauf, puis retourner sur la Terre, tout cela n'est qu'un rêve fou digne de Jules Verne.»

Lee de Forest, ingénieur, inventeur de la triode en 1907, 1957.

«Les seuls domaines d'intérêt pour la délégation à l'informatique sont la gestion et le calcul scientifique.»

Maurice Allègre, délégué à l'informatique, en réponse à une question de l'auteur sur l'informatique industrielle, 1969.

«Il n'y a aucune raison pour qu'un individu quelconque possède un ordinateur chez lui.»

Kenneth Olsen, cofondateur et président de DEC, père des PDP, 1977.

«On peut refermer le livre des maladies infectieuses»

Le médecin général des services de santé américains, 1969.

Depuis, nous avons eu le virus Ébola, les divers VIH et les prions.

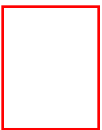
«Je crois qu'OS/2 est destiné à être le système d'exploitation le plus important de tous les temps.»

Bill GATES, PDG et fondateur de Microsoft, 1988.

«L'époque des PC est terminée.»

Lou Gerstner, Directeur d'IBM, 1998.

Lecteur, n'hésite donc pas à dire des bêtises, tu seras en bonne compagnie.



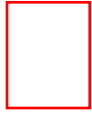
Conservatoire national des arts et métiers
Architectures des systèmes informatiques
ANNEXE 07
Réflexions éparses et bêtisier
Année 2002-2003

Conservatoire national des arts et métiers
Architectures des systèmes informatiques

ANNEXE 08

Microprocesseurs, Intel (beaucoup) et Motorola (un peu)

Année 2002-2003



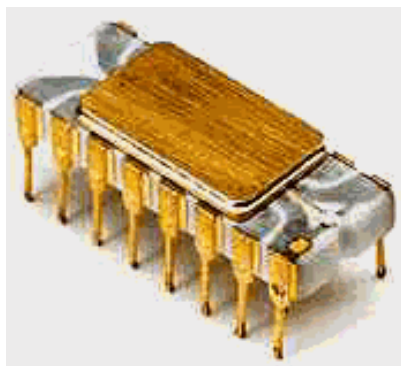
Cette annexe contient une vue rétrospective non exhaustive des gammes des deux fabricants dont l'activité a été continue depuis le début des microprocesseurs. Les Mips mentionnés ici et là, sont ceux fournis par le constructeur.

Le microprocesseur, ensemble des circuits essentiels d'un processeur groupés sur un circuit intégré unique, a été imaginé par Gordon E. Moore, qui quitta le laboratoire Livermore pour fonder la société Intel (Integrated Electronic) avec Fred Noyce en 1968.

Intel Corporation, réalisait des circuits intégrés à la demande. Le microprocesseur à usage (presque) général apparaît en 1971. Il existait alors des microprocesseurs pour l'exécution de tâches très spécialisées.

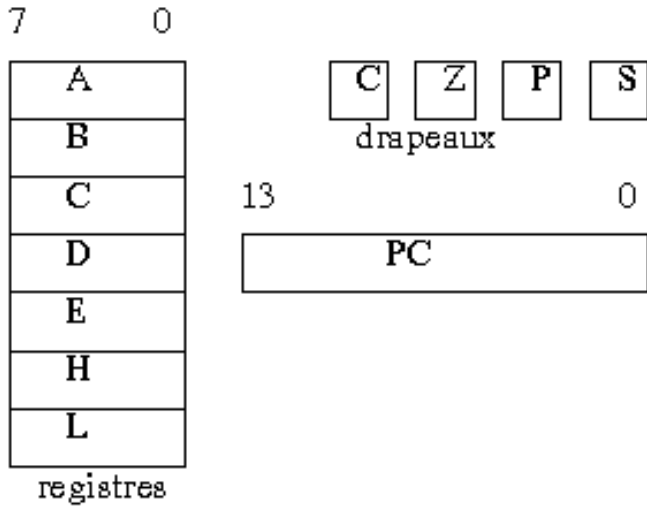
1971 **4004**, registres de 4 bits, technique Pmos, gravure à 10 μm , 108 kHz, 2250 transistors, 46 instructions, boîtier à 16 broches, 1 k de mémoire de données, 4 k de mémoire d'instructions, 16 registres de 8 bits, 6000 instructions par seconde, 200 dollars l'unité, capacité d'adressage 1 ko. Il est aussi puissant que l'ENIAC. Il fut conçu à la demande de Busicom, fabricant japonais d'horlogerie.

Son format de base était l'arithmétique décimale codée binaire DCB. Ce premier microprocesseur commercial a pour auteurs Ted Hoff, Federico Faggin et Stan Mazor. Intel annonce le premier microcalculateur MCS-4 avec un 4004, une puce ROM 4001, une puce RAM 4002, et une puce de registre à décalage 4003. Il eut une version améliorée, le 4040. Cependant, l'idée de créer un ordinateur accessible à tous n'est pas encore née.



1972 **8008** registres de 8 bits, adressage de 16ko (CO sur 14 bits), 3300 transistors, technique Pmos 10 μm à 200 kHz, 6 000 instructions par seconde, boîtier à 18 broches, conçu à la demande de Computer Terminal Corporation (plus tard DataPoint) pour commander les écrans cathodiques que fabriquait cette entreprise. Son jeu d'opérations était différent du précédent car il était conçu pour manipuler des octets et non plus des nombres. Intel avait ajouté quelques

instructions aux spécifications du client pour en faire un processeur à usages multiples. Il faisait des calculs booléens sur des octets. Il avait 4 drapeaux d'état : retenue (complément) C, zéro Z, parité P, signe S, un accumulateur A, six registres de 8 bits B C D E H L, une pile matérielle pour 8 adresses de retour.

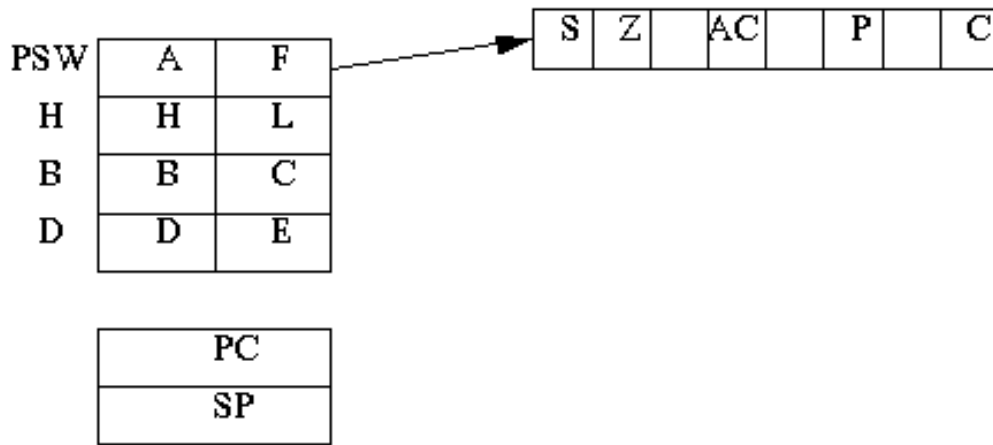


Les 66 codes d'instructions le rendaient proche des plus anciens ordinateurs (la CAB500 en avait 48). L'adressage était soit immédiat soit indexé, l'index était la concaténation de 6 bits du registre H et de 8 bits du registre L.

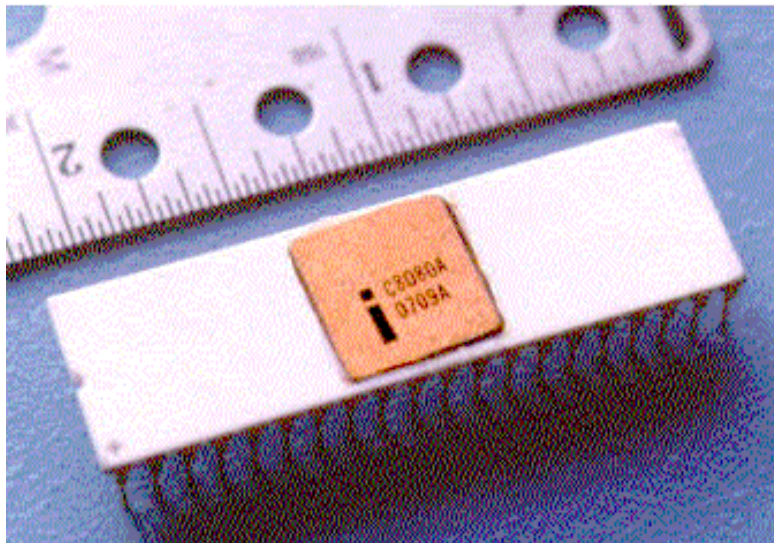
Les partis qui avaient été pris avaient pour première cause les 18 broches des boîtiers standard, ainsi le bus d'adresse de 8 bits de largeur était multiplexé, poids faible d'abord, poids fort ensuite, provoquant la recherche simultanée de l'octet dans tous les boîtiers mémoires de 256 octets, le boîtier était choisi ensuite par la partie haute de l'adresse.

Il était le processeur du Micral 8 (mai 1973), premier microordinateur intégré car il n'est pas un kit à monter, conçu au départ pour la régulation par François Gernelle et lancé par Truong Trong Thi, qui avait fondé et dirigeait la société Réalisations électriques et électroniques REE.

1974 **8080**, registres et bus de données de 8 bits, adressage de 64 ko, 4500 transistors, technique Nmos 6 µm à 200 kHz, 64000 instructions par seconde (0,064 Mips), boîtier à 40 broches. La compatibilité est totale avec le précédent. Le boîtier de 40 broches simplifie le bus. L'adressage sur 16 bits donne 64 ko de mémoire centrale possible. Il est accompagné du contrôleur d'entrées-sorties et d'interruptions 8228 et d'une horloge 8224. Quelques instructions font intervenir des adresses de 16 bits par l'utilisation de couples de registres BC, DE et HL. L'adressage direct devient possible. Il a une addition sur 16 bits mais ni multiplication ni division. La pile en mémoire a un pointeur de 16 bits. Mais toutes ces extensions ne sont pas symétriques (cela aurait demandé beaucoup plus de transistors). Le mécanisme d'interruptions est original, le périphérique fournit un code opération (cf. § 2.9.2.1).



Le 8080 devient rapidement un standard de l'industrie.



1976 **8085**, registres de 8 bits, adressage de 64ko, 6500 transistors, technique Nmos 3 μ m à 5 MHz, 0,37 Mips. Il a besoin d'une seule alimentation de 5 volt. C'est un 8080 qui intègre le contrôleur d'interruptions (8228) et d'entrées et sorties série et le circuit d'horloge (8224). Deux instructions ont été ajoutées pour accéder au masque d'interruption et pour gérer la voie série. Certains auteurs ont considéré le 8085 comme un microcontrôleur malgré son succès comme unité centrale de micro ordinateurs. D'autres microcontrôleurs ont été fondés sur le 8080, avec parfois de la mémoire morte ou des entrées-sorties plus complexes; par exemple, le 8048 des claviers.

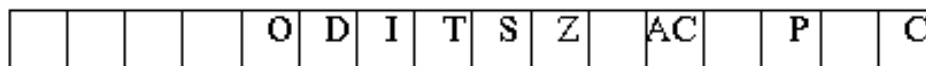
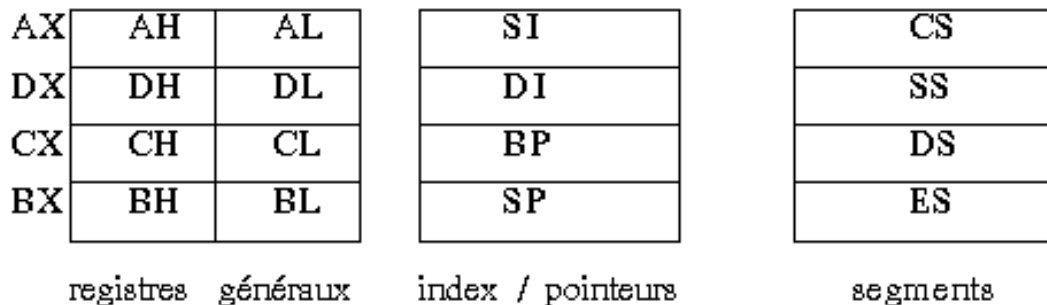
Pour la première fois on voit apparaître des processeurs compatibles concurrents : le plus connu est le Z80, sorti en juillet 1976, de la société Zilog fondée par des anciens d'Intel. Il a un jeu d'instructions plus complet : manipulations de bits et de chaînes de caractères notamment et un dispositif intégré de rafraîchissement de la mémoire, au total 80 instructions de plus que le 8085. Il contient 8500 transistors et fonctionne à 2.5 MHz. Des versions améliorées sortiront plus tard : Z-80B, Z-180, Z-181 et Z-182.

1978 **8086**, registres de 16 bits, adressage de 1 Mo, 29000 transistors, technique Cmos faible consommation, 3 μ m à 4,77 MHz, 0,33 Mips, boîtier à 40 broches. En mai 1978, la production commence, il est annoncé en juin 1978 au prix de 360 dollars. Il sera porté ultérieurement à 8 MHz (0,66 Mips) et 10 MHz.

Selon les auteurs, il a été conçu en trois ans pour les uns, en trois semaines par deux ingénieurs pour les autres !, quand le i432 a été manifestement un échec.

Intel n'était pas le premier à sortir un microprocesseur 16 bits : en 76 étaient apparus des processeurs peu diffusés, le TMS9900 de Texas Instrument, le 9440 successeur du F8 de Fairchild, le LSI11 de DEC, l'IMP 16 processeur en tranches de 4 bits de National Semiconductor.

La compatibilité binaire avec le 8080 disparaît mais quelques moyens sont fournis pour récupérer les sources. L'arithmétique est signée sur 8 et 16 bits, les chaînes sont gérées, les programmes peuvent être relogeables et réentrants. tout en multipliant les performances par un facteur 10. Les registres sont toujours spécialisés bien que les accumulations soient faisables dans tous au lieu du seul accumulateur : BX est un pointeur de base en mémoire, CX est un compteur, DX est un pointeur de données et AX reste seul accumulateur pour certaines instructions (multiplication, division, entrées-sorties). SI et DI sont des index source et destination, SP est le pointeur de pile et BP est le pointeur de base sur la pile. Les modes d'adressages deviennent puissants (possibilité d'ajouter deux registres base et index, et un déplacement) mais lents.



La segmentation dispose de quatre segments par processus. Chaque segment est limité à 64ko et nécessite des modèles de compilation pour les langages évolués. Les registres de segmentation sont difficiles à manipuler sans passer par la mémoire. Les instructions de mouvement ne positionnent pas de drapeaux. La virgule flottante est traitée par un circuit optionnel. Les 40 broches obligent à des acrobaties, plusieurs broches sont multiplexées. Par exemple, une broche donne le choix entre deux modes de fonctionnement : minimum et maximum. Le mode minimum offre un bus d'adresses de 16 bits sans prendre en compte le coprocesseur optionnel (8087), le 8086 étant alors un super-8080. Le mode maximum demande un gestionnaire de bus et gère le coprocesseur, la capacité d'adressage est alors 1 Mo via les segments de 64 Ko mais le bus d'adresse est multiplexé.

Le mécanisme d'interruption est à 256 niveaux, l'espace d'adressage des entrées-sorties passe à 64 Ko.

Une ébauche de pipeline apparaît avec deux unités opérant simultanément : l'Execution Unit et la Bus Interface Unit qui réalise les accès à la mémoire et anticipe le chargement des instructions. Les gain apporté est de l'ordre de 10%.

Son premier clone sera le V30 de NEC à 8 MHz avec 63000 en 1984.

La même année, Intel produit le contrôleur d'interruptions 8259A.

1979 Seattle Computer Products produit en milieu d'année, la première carte mère basée sur un 8086 et un bus S100.

1979 **8088**, registres de 16 bits, bus de données de 8 bits, adressage de 1 Mo, 29000 transistors, technique Cmos faible consommation, 3 μ m à 4,77 MHz, 0,33 Mips et 40 broches. Il sera porté ultérieurement à 8 MHz (0,75 Mips). C'est une restriction du précédent avec un bus de données multiplexé sur 8 bits, processeur de l'*IBM PC*. Son premier clone est le V20 de Nec à 8 MHz, produit en 1984, avec 63000 transistors. Le boîtier avait 40 pattes, 20 d'adresses, 8 de données, 19 de commande, une d'horloge, deux de terre, 8 pattes de données et 4 de commandes étaient multiplexées avec 12 pattes d'adresses.

1980 **8087**, C'est le coprocesseur mathématique des 8086 et 8088. La même année, Intel annonce l'iAPX 432, futur processeur 32 bits.

1981 **iAPX432** est présenté au printemps par Intel pour une performance attendue de 2 Mips. Il sera produit en 1982, il gère un espace virtuel de 1 Téra octets. Son jeu d'instructions est conçu pour les langages de haut niveau, tel ADA. Il n'est pas compatible avec la famille x86 et disparaît rapidement.

1981 août IBM annonce l'IBM 5150 PC Personal Computer à New York. Il utilise le 8088, a 64 Ko de RAM, 40 Ko de ROM, une disquette de 5 pouces $\frac{1}{4}$ de 160 Ko, son système d'exploitation est PC-DOS 1.0. Il coûte 3000 dollars en noir et blanc, 6000 dollars avec un moniteur CGA, 16 couleurs. La même année, le projet iAPX432 est arrêté par Intel.

1982 **80186**, est une extension du 8086 par intégration sur la puce de ressources supplémentaires d'entrées et sorties : un canal d'accès direct, un contrôleur d'interruptions, une horloge, des compteurs/temporisateurs et un décodage d'adresse pour sélectionner les boîtiers de mémoires. Il a dix instructions nouvelles. Il a eu une vie de micro contrôleur sauf dans le TANDY TRS 2000.

1982 **80188**, est l'homologue du 80186 vis-à-vis du 8088.

1984 **80286**, registres de 16 bits, bus de données sur 16 bits, adressage de 16 Mo, 1 Go de mémoire virtuelle, 134000 transistors, technique Cmos faible consommation, 1,5 μ m à 6 MHz, 0,9 Mips, son boîtier a 68 broches. En mai 1978, la production commence, il est annoncé au prix de 360 dollars l'unité par quantités de 100. Il sera porté ultérieurement à 10 MHz (1,5 Mips) et 12 MHz (2,66 Mips).

Il peut gérer une mémoire virtuelle segmentée. Il n'y a plus de multiplexage sur le bus. Il a des comportements différents selon deux modes : dans le mode réel il est compatible binaire avec le 8086, dans le mode protégé, ses segments deviennent des sélecteurs dans des tables de descripteurs (jusqu'à 8192 par table), avec une capacité d'adressage de 1 Go pour chaque tâche. Les tâches sont gérées directement par le processeur, ainsi que les protections des segments avec 4 niveaux de privilèges. Encore faut-il que les systèmes d'exploitation utilisent ces ressources ce qui n'est pas le cas du DOS de Microsoft qui est à ses versions 3 sur les 286.

Les fonctions de pipeline sont améliorées, quatre unités opèrent simultanément : l'interface avec le bus (bus unit), le décodeur d'instructions (Instruction Unit), l'unité d'exécution (Execution Unit) et le générateur d'adresses physiques (Address Unit).

Il est le processeur de l'IBM AT (advanced technology).

1985 **80287** est le coprocesseur arithmétique du 80286.

1985 **80386** est l'extension du 80286, registres, bus, données et adresses sur 32 bits. Intel l'annonce à 16 MHz version 80386DX. Il est composé de 275,000 transistors, technique 1.5 µm. Son prix initial est de 299 dollars. Il adresse 4 Go de mémoire physique et jusqu'à 64 To de mémoire virtuelle. Il reste compatible avec le 8086 dans son mode réel. Il est compatible dans son mode protégé avec le 286. Il exécute ainsi des tâches 16 bits (286) que des tâches 32 bits (386). Un autre type de tâche est par ailleurs introduit : les tâches virtuelles 86, pour exécuter les programmes du 8086 en mode protégé.

Le pipeline est encore augmenté, la plupart des instructions sont exécutées en 2 à 5 cycles d'horloge (suivant qu'un accès mémoire est nécessaire ou non).

Un mécanisme de pagination est intégré au processeur : l'unité de gestion de la mémoire segmente et de pagine. Pour une tâche, l'espace d'adressage virtuel passe à 64 Téra-octets, dans un espace mémoire physique possible de 4 Go.

De nouvelles instructions assez nombreuses manipulent les bits, font des décalages, prennent en compte des facteurs additifs pour la manipulation des tableaux.

La même observation que plus haut sur les systèmes d'exploitation reste valable. Unix est assez aisément disponible sur ces machines.

À la fin de 1986, Zilog sort sans grand succès le Z-80000. Il avait des segments d'adressage de 24 bits, un cache de 256 octets, il fonctionnait en multiprocesseur et avait un pipe line à 6 étages.

1987 **80386DX** à 20 MHz et **80387**, coprocesseur mathématique de 80386.

1988 **80386DX** à 25 MHz (5,5 Mips)

80386SX à 16 Mhz (2,5 Mips), son bus de données est réduit à 16 bits. Il coûte 219 dollars l'unité par quantité de 100.

860XR processeur quasi Risc en CHMos IV à 1 µm

1989 **80386SX** à 20 Mhz (2,5 Mips).

860CA processeur superscalaire exécute deux instructions simultanément.

80386DX à 33 MHz et **80387** refondus.

80486 annoncé et mis en production à 25 MHz (20 Mips) et au prix de 900 dollars. Il est une extension du 80286. Il contient 1,2 millions de transistors, technologie 1 µm. L'opérateur de calcul en virgule flottante jusqu'alors externe et optionnel (8087, 80287, 80387) est intégré dans la puce. Il devient un opérateur qui un accès immédiat aux opérandes. Une nouveauté (chez Intel) apparaît, un cache interne banalisé de 8 Ko. Le pipeline est encore plus détaillé, le décodage est fait en deux étapes, le pipeline a 5 étages : prefetch , decode1, decode2,

execute, write back).

Avec une optimisation de l'ordre des instructions et le cache interne, les instructions les plus courantes s'exécutent en un cycle d'horloge dans le meilleur cas.

1990 **80387SX** nouveau coprocesseur.

80486 à 33 MHz (27 Mips).

80386SL, processeur pour portables à 20 MHz (4,2 Mips), 855000 transistors, technique à 1 µm, registres de 32 bits, bus de 16 bits, à 150 dollars l'unité par quantité de 1000.

Nouvelles versions des 80387 à 16, 20 et 25 MHz. Une décision de justice signifie à Intel que les numéros tels que X86 sont utilisables par toute autre entreprise. Intel décide de nommer ses futurs processeurs différemment.

1991 **80486SX** à 20 MHz (16,5 Mips), c'est un 486 sans le coprocesseur, prix 258 dollars.

80486 à 50 MHz (41 Mips), technologie 0,8 µm.

80486 à 25 MHz (20 Mips), technologie 0,8 µm.

80486SX à 13 MHz (16 Mips).

80386SL à 25 MHz (5,3 Mips).

Début de l'étude du successeur du Pentium

1992 Nouvelles versions des 80386 : DX2 à 33/66 MHz et SX à 33 MHz

82489DX contrôleur d'interruptions

80486DX2 41 Mips, horloge externe 25 MHz, horloge interne 50-MHz. C'est pour l'essentiel un 25 MHz dont la fréquence interne est doublée.

1993 Spécifications définitives du bus PCI

Pentium. En mars 1993, Intel abandonne la numérotation de ses processeurs. Le motif est juridique car des nombres ne peuvent pas être déposés comme noms de modèles. Il annonce le Pentium, technique BiCMos à 0,8 µm. Il opère sur 32 bits. Pour la première fois, l'essentiel de la puce est consacré à des caches pour les instructions et les données séparément et à des unités d'exécution en pipeline. Le microcode n'occupe plus que 3% de la surface.

Le cache de premier niveau est passé de 8 à 16 Ko. Il est organisé en lignes de 32 octets, deux fois plus que dans le 486 et l'interface de bus est capable de remplir une ligne avec une simple lecture en rafale de quatre morceaux de 64 bits. Le cache des données pratique le «writeback». Un bus de 256 bits sort du cache instruction pour remplir les tampons du «prefetch» (32 octets) en un seul cycle. Le bus d'adresse et les registres restent sur 32 bits.

Le bus de données du processeur a été étendu à 64 bits.

Le processeur a deux unités de calcul entier opérant simultanément. Elles sont alimentées chacune par un pipeline d'instructions, il est donc superscalaire. Les conditions de fonctionnement simultanées sont : deux instructions de calcul entier consécutives, pas de dépendance de données entre elles, pas de déplacement ni de valeur immédiate, pas de préfixe, et enfin ne manipulent pas la pile ni ne fassent de comparaison.

L'unité de virgule flottante est alimentée par un pipe line à huit étages, dont cinq sont communs avec les deux pipelines entiers : l'étage write back des unités entières est ici remplacé par X1 (conversion des nombres flottants et écriture dans les registres du FPU), puis on trouve X2 (exécution des opérations arithmétiques), WF (arrondi et writeback) et enfin ER (rapport d'erreurs et mise à jour du statut). Pour pouvoir charger un opérande flottant double précision (64 bits) en un seul cycle, les deux unités entières travaillent de concert (on ne peut donc pas avoir en même temps instructions entières et flottantes). Les deux étages d'exécution ne peuvent pas travailler sur deux instructions différentes à une exception près : l'instruction FXCH (qui échange le sommet de la pile flottante avec un élément plus profond) peut être couplée avec les opérations arithmétiques simples (FADD, FSUB, FMUL, FDIV...)

En moyenne, le Pentium 75, première entrée de gamme est deux à trois fois plus rapide que le prédécesseur 486 DX66. C'est la première fois que le gain apporté par un processeur nouveau est aussi faible, alors que :

- la quantité de mémoire cache interne est doublée;
- le bus de données passe de 32 à 64 bits;
- l'unité de calcul en virgule flottante, entièrement redessinée devient cinq fois plus rapide;
- un second pipeline, c'est-à-dire la file d'attente de décodage et d'exécution des instructions, est intégré;
- un circuit de prédiction de branchement apparaît;
- il contient 3,1 millions de transistors contre 1,2 million dans le 486 DX2-66.

1994 **80486SX** à 16 MHz (13 Mips).

1994 **Pentium** livraison des 60/90 MHz (150 Mips), technologie 0,6 µm BiCMos, 849 dollars l'unité par quantité de 1000.

Pentium livraison des 66/100 MHz (166 Mips), technologie 0,6 µm BiCMos, 995 dollars l'unité par quantité de 1000.

Pentium livraison des 75 MHz (127 Mips), technologie 0,6 µm BiCMos, 3,2 million de transistors.

1995 **Pentium** 120 MHz (203 Mips), 935 dollars l'unité par quantité de 1000.

Pentium 133 MHz, technologie 0,35 µm, 3,2 million de transistors, 219 Mips, 935 dollars l'unité par quantité de 1000.

Juin 1995, arrêt des livraisons des Pentium 60 et 66 MHz.

1996 **Pentium Pro**. Il est plus rapide que le précédent. Le Pentium Pro a 5,5 millions de transistors. Il est la sixième génération des processeurs 80x86. Son architecture est nommée P6. Il accentue l'orientation prise par le Pentium en direction des techniques des processeurs

Risc, il est lui aussi superscalaire, il a de plus ce que Intel appelle l'exécution dynamique des instructions, la prédiction dynamique des branchements, l'exécution non séquentielle et l'exécution spéculative.

Les instructions de la famille 80x86 sont exécutées non plus par un microprogramme ad'hoc mais par une suite d'opérations simples de type Risc, qu'Intel appelle micro-opérations. Ces instructions Risc sont ensuite distribuées aux unités d'exécution: chargement, déchargement, flottante et entières. Intel nomme cela exécution dynamique. Le processeur peut ainsi être vu en deux couches, une couche basse Risc et une couche haute Cisc.

Un cache secondaire nouveau de 256 Ko de mémoire statique (15,5 millions de transistors) est relié à l'unité centrale par un bus spécial de 64 bits à la même fréquence que celle-ci. Les Pentium à 90 ou 120 MHz ont un bus externe à 60 MHz. Ce bus supplémentaire évite tout conflit sur le ou les autres bus pour l'accès à ce cache.

Le pipeline du Pentium Pro a 14 étages. Il combine décodage des instructions Cisc, appel des instructions Risc et ordonnancement de celles-ci.

1996 **Pentium MMX**, 4,5 millions de transistors. Il contient un jeu supplémentaire d'instructions (§ 2.2.1) pour le traitement des objets visuels et sonores par une technique de SIMD. Il s'agit de traiter simultanément à l'identique, plusieurs données (4 au plus). Le cache de premier niveau est porté à 32 Ko.

1997 **Pentium II**, 7,5 millions de transistors. L'architecture P6 est maintenue. La fréquence maximale des produits commerciaux courant 1998 est de 450 MHz.

2000 **Pentium IV** est profondément renouvelé avec :

- un bus à 400 MHz sur 128 bits;
- 20 ko de cache L1 et 256 ko de cache L2, passé à 512 ko en 2000;
- un cache nouveau de 12 ko qui contient le résultat de la transformation des dernières instructions x86 en instructions RISC;
- une UAL fonctionnant à fréquence double du reste de la puce;
- des fréquences commençant à 1500 MHz, actuellement 2,8 MHz;
- un pipeline à 20 étages;
- 42 millions of transistors;
- de la mémoire Dual Rambus.

2001 **Itanium** est le premier d'une famille de processeurs à données sur 64 bits.

On en trouvera les caractéristiques principales à :

<http://www.intel.com/design/itanium/itanium/ItaniumProdBrief.pdf>

Les objectifs propres à cette nouvelle génération paraissent être :

Simplicité

En revenant aux conceptions initiales des RISC.

Extensibilité

Il contient 128 registres généraux et 128 registres pour nombres flottants, ce qui apporte des capacités de type superscalaire importantes.

Parallélisme et adaptation aux résultats de la compilation par la technique dite EPIC «explicit

parallel instruction computing».

Calcul sur 64 bits

Il s'agit d'un vrai 64 bits autant pour les types de données que pour les entiers longs.

Une capacité importante de mémoire principale.

Un espace virtuel considérable.

Nom	Date	Horloge MHz	Bus bits	Transistors Gravure	Mémoire réelle adressable	Mémoire virtuelle adressable
4004	Nov 71	0,108	4	2300 10 μm	640 octets	
8008	Jan 72	0,108	8	3500	16 Ko	
8080	Jan 74	2	8	6000 6 μm	64 Ko	
8086	Août 78	5 puis 8, 10	16	29000 3 μm	1 Mo	
8088	Juin 79	5 puis 8	8	29000 3 μm	1 Mo	
80286	Fév 82	8 puis 10, 12	16	134000 1,5 μm	16 Mo	1 Go
80386DX	Oct 85	16 puis 20, 25, 33	32	275000 1 μm	4 Go	64 To
80386SX	Juin 88	16 puis 20	16	275000 1 μm	4 Go	64 To
80486DX	Avr 89	25 puis 33,50	32	1,2 million 1 μm puis 0,8 μm à 50 MHz	4 Go	64 To
80486SX	Avr 91	16 puis 20,25,33	32	1,185 million 0,8 μm	4 Go	64 To
Pentium	Mars 93	60 puis 66, 75, 90, 100, 130, 133, 150, 166	32	3,1 million 0,8 μm	4 Go	64 To
Pentium Pro	Mars 95	150 puis 180, 200	64	5,5 million 0,32 μm	4 Go	64 To
Pentium IV	2000	1500 à 2800	64	42 millions 0,15 μm puis 0,13 μm	4 Go	64 To

Performances relatives des processeurs 386 Intel

L'indice iComp est construit avec des résultats de bancs d'essais qu'Intel a utilisé pour évaluer ses processeurs. L'indice de référence 100 est donné au processeur 486 SX à 25 MHz (un 486 sans coprocesseur arithmétique intégré), qui était donné pour 20 Mips. Les autres indices sont calculés par rapport à lui.

Pentium/66 : 565
 486 DX2/66 : 297
 486 DX/50 : 249
 486 DX2/50 : 231
 486 DX/33 : 166
 486 SX/33 : 136
 486 DX/25 : 122
 486 SX/25 : 100
 486 SX/20 : 78
 386 DX/33 : 68
 386 DX/25 : 49
 386 SX/25 : 39
 386 SX/20 : 32

Motorola Corporation, à la différence d'Intel, n'a pas pour seule activité les circuits intégrés, processeurs, mémoires et circuits annexes. Cette société est très présente dans d'autres domaines de l'électronique, les téléphones notamment.

Les processeurs de Motorola ont été :

Nom	Année	Registres	Bus Données	Bus Adresses	Commentaires
6800	1974	8b	8b	64 Ko	1 registre de données, 2 registres d'index, 1 pointeur 4000 transistors, une alim. 5V
6809	1976	8b	8b	64 Ko	arithmétique 16 bits, 2 MHz.
68000	1979	32b	16b	16 Mo	2 Mips
68008	1982	32b	8b	4 Mo	bas de gamme (cf. 8088 Intel)
68010	1983	32	16	16 Mo	mémoire virtuelle
68012	1983	32	16	2 Go	68010 étendu
68020	1984	32	32	4 Go	UC 32 bits, 200000 transistors, techn. CMOS 1,7 µm, 8MHz, 12,5 MHz, 16,7 MHz, 20 MHz, 17 registres, pipeline, cache 256 octets, pagination.
68030	1987	32	32	4 Go	300000 transistors, UGM intégrée.

88000	1988				annonce de ce Risc, sans suite.
68040	1990	32	32	4 Go	1200000 transistors, 25 MHz puis 50 MHz, 20 MIPS, unité flottante, caches de Harvard, disponible en 1991.
68060	1994	32	32	4 Go	2,5 millions de transistors, techn. CMOS 0.5µm? métallisation 3 couches, 223 broches, superscalaire à deux unités.

Motorola ne fabrique plus de microprocesseurs nouveaux exclusifs.

La gamme PowerPC est commune à Motorola sous les noms MPC6xx (MotorolaPC) et à IBM sous les noms PPC6xx (PowerPC).

La concurrence en 1999.

Les producteurs de microprocesseurs sont aujourd'hui nombreux :

- **AMD** : Am486, Nx586, Am5x86, K5, K6, K6-2.
- **Cyrix** : MediaGX, 6x86MX, M2.
- **Digital** : Alpha 21164, 21264.
- **Fujitsu** : Sparc64.
- **Hewlett Packard** : PA7300, PA8000, PA8200, PA8500, PA8700.
- **Hitachi** : Famille SuperH SH7750.
- **IBM** :

gamme PowerPC : 601, 602, 603, 604, 750 et RS64,

gamme Power : Power2, Power3,

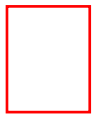
gamme x86 : sur plans de Cyrix : 5x86, 6x86, 6x86MX.

- **Intel** : i960, 80960RP, Pentium, Pentium overdrive, Pentium MMX, Pentium Pro, Pentium II, Celeron, Xeon;
- **Mips** : R3000, R4000, R5000, R8000, R10000, R12000;
- **Motorola** : PowerPC 603, 740, 750 (G3), 620, G4;
- **Nec** : VR5000, VR5400, VR10000;
- **Quantum effect design** : RM7000;
- **STMicroelectronics** (ex SGS Thomson) : STPC Consumer;
- **Sun** : Microsparc II et Ultrasparc I, II et III.

Signalons enfin des producteurs de processeurs de type x86 à bas prix :

- **IDT** et notamment IDT C6 qui inclut la technique MMX;
- **VIA Technologies** produit le C3 incluant MMX, 3DNow! avec une gravure à 16 µm;
- **Rise Technology**;
- **Metaflow** (acheté par STMicroelectronics);
- **Transmeta**, processeurs Crusoe.

Texas Instruments ne produit que des processeurs de traitement du signal .



Conservatoire national des arts et métiers
Architectures des systèmes informatiques
ANNEXE 08
Microprocesseurs, Intel (beaucoup) et Motorola (un peu)
Année 2002-2003

Conservatoire national des arts et métiers
Architectures des systèmes informatiques
 ANNEXE 09
Comparaison de processeurs de 1998
 Année 2002-2003



1 COMPARAISON DE PROCESSEURS CISC

Les processeurs examinés dans cette annexe sont les :

- Pentium II d'Intel;
- K6 modèle 7 d'American Micro Device;
- 6x86MX de Cyrix du groupe National Semiconductors.

Ils ont été choisis plutôt que les processeurs Sparc ou Mips pour leur grande diffusion. Le PowerPc fait l'objet d'une annexe pour lui seul.

Ces trois processeurs sont compatibles au sens de leur [visibilité par le programmeur](#). On en attend des caractères communs. Ces ressemblances sont encore accentuées par la similarité de leur technique de fabrication.

La comparaison est faite dans le tableau ci-dessous, les commentaires seront faits à la suite.

Caractère	Intel Pentium II	AMD K6 modèle 7	Cyrix 6x86MX
horloge interne MHz	266 à 450	266, 300	133 à 250
horloge bus MHz	66 ou 100	66 ou 100	66, 75, 83, 100
architecture	x86 --> Risc	x86 --> Risc	x86 directe
multimédia MMX	oui, 57 instructions	oui, 59 instructions	oui, 57 instructions
virgule flottante	oui, bus 32,64, 80 bits	oui, bus 64 bits	oui, bus 64 bits
pipelines	un ramifié : 9 étages communs + 3 étages parallèles	sept 6 étages chacun	deux 7 étages chacun 1 seul pour MMX, FPU
décodage x86 à Risc86	3 décodeurs, 6 instructions au plus décodées par cycle	4 décodeurs dont 1 pour MMX 2 pour FPU 2 instructions décodées par cycle et par décodeur	sans objet, le coeur n'est pas Risc
ordonnanceur	analyse de dépendance	examen de 12 instructions tampon de 24 opérations envoi de 6 opérations	dépendances traitées par renommage des registres
exécution parallèle	2 entières 2 FPU 1 accès mémoire + 1 MMX	entière X entière Y chargement stockage FPU branchement + 1 MMX	deux UAL

Annexe 9, Comparaison de processeurs

cache interne	1 ^{er} niveau L1 : 16 Ko instructions 16 Ko données 2 ^d niveau L2 fréquence moitié de l'horloge 512 Ko indifférencié jusqu'à 2 Mo pour les P à haute fréquence	1 ^{er} niveau L1 seul: 32 Ko instructions avec prédécodage 32 Ko données	1 ^{er} niveau L1 seul : 64 Ko unique à 4 accès simultanés + ligne d'instructions de 256 octets
prédiction de branchement	2 niveaux cache cible à 512 entrées	2 niveaux : 16 entrées + table statistique à 8192 entrées calcul à la volée	cache cible de 512 entrées à 4 accès + table statistique à 1024 entrées
taux annoncé de bonnes prédictions	90 %	90 à 95 %	90 %
renommage des registres	oui, 40 registres	oui, 48 registres	oui, 32 registres
tampon TLB	non ?	64 entrées instructions 128 entrées données	1 ^{er} niveau 16 entrées 2 ^e niveau 384 entrées
exécution spéculative	oui	oui	oui
gravure		0,35 microns 5 couches	
nombre de transistors		8,8 millions	

RESSEMBLANCES ET DISSEMBLANCES

Les ressemblances et dissemblances vont porter sur les moyens mis en oeuvre.

1) Le moyen d'exécution des instructions.

On a longtemps distingué les machines CISC des machines Risc. Récemment les processeurs mixtes sont apparus. Le Pentium II et le K6 en font partie. Cette mixité est construite comme suit :

Le jeu d'instructions du programmeur est CISC, ici x86 pour les trois processeurs.

- L'exécution dans le Cyrix utilise les moyens habituels, partie microprogrammée, partie câblée.
- Dans l'AMD K6, la technique est celle de l'émulation. On émule la machine x86 sur une machine Risc. La première réalisation de cette technique est due à NexGen en 1994. La machine de base est Risc, elle possède son jeu d'instructions non visibles de l'utilisateur. Chaque instruction Cisc est décodée et remplacée par la suite d'instructions Risc qui va réaliser son exécution. La machine Risc de base joue le rôle de la micromachine de la microprogrammation. En ce sens la suite des instructions Risc est analogue à la suite des microinstructions. La différence est que le contenu d'une microinstruction est une suite de microcommandes, alors qu'ici les instructions Risc sont traitées par un séquenceur câblé normal.
- La situation du Pentium est intermédiaire. Les instructions simples sont décodées directement comme étant des instructions RISC.

2) Les instructions MMX.

Elles sont quasiment communes aux trois processeurs.

3) La virgule flottante

Une unité spécialisée existe dans les trois processeurs.

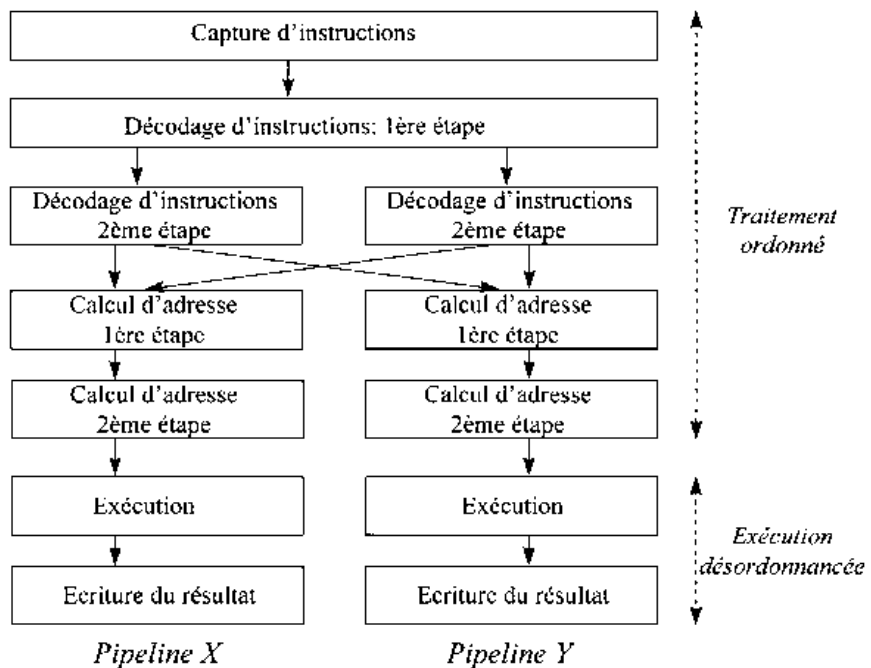
4) L'alimentation en instructions

Elle est faite par pipe line dans les trois processeurs avec des différences notables :

Le Cyrix possède deux pipelines alimentés par un tronc commun de deux étages comme sur la figure ci-dessous.

La capture est faite par 16 octets à la fois à partir du cache. La première étape de décodage marque les instructions en longueur et en transmet au plus deux, une à chaque étage de décodage deuxième étape. Les étages ID2 alimentent indifféremment les deux caches selon l'occupation de ceux-ci. La notion de pipe line n'en est que légèrement brouillée.

La suite des opérations est classique, y compris l'exécution dans l'une des deux unités de traitement et l'écriture du résultat soit dans un registre soit dans le tampon du cache.



Le K6 a 7 unités d'exécution spécialisées, les 6 premières exécutent des opérations Risc86 : (voir la figure suivante)

```

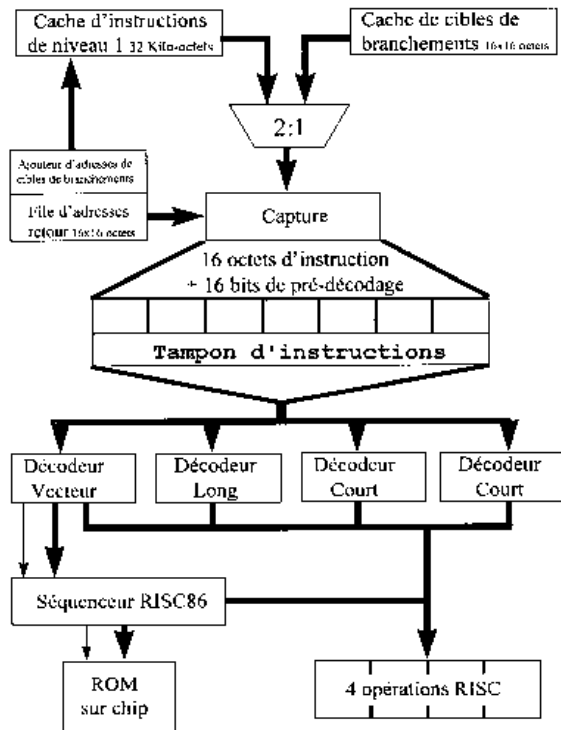
.chargement;
.stockage;
.calculs entiers X;
.calculs entiers Y;
.calculs multimédia;
.virgule flottante;
.branchement.
  
```

La présence des deux unités de chargement et de stockage est révélatrice de la machine Risc sous-jacente. Cette machine exécute le jeu d'instructions Enhanced Risc86 ou Risc86, repris des réalisations de la société NexGen achetée par AMD.

La notion de pipeline est ici fortement brouillée. On peut bien sûr considérer qu'il y en a 7 puisque sept unités d'exécution sont alimentées. On peut aussi dire qu'il y en a 4, selon le nombre de décodeurs. De toutes façons les pipelines passent tous par le dispositif unique nommé tampon d'instructions. Nous sommes loin de l'analogie du tuyau unique menant d'une extrémité à l'autre.

```

.les bits de pré décodage marquent les limites des instructions x86;
.les décodeurs courts traitent les instructions x86 dont la traduction se fait en zéro, une ou deux instructions Risc86;
.le décodeur long traite les instructions x86 dont la traduction se fait en quatre instructions Risc86 ou moins;
.le décodeur dit vecteur traite les instructions x86 plus complexes.
  
```



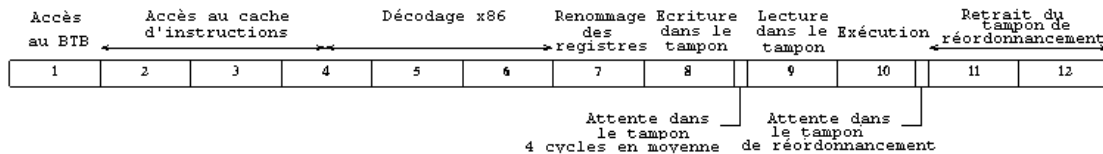
Tampon d'instructions et décodage

Le Pentium II a un pipeline unique à 9 étages, auquel sont ajoutés trois étages de decodeurs en parallèle.

Ces 12 étages sont groupés en 4 blocs :

- .capture et décodage;
- .pool d'instructions;
- .affectation et exécution;
- .retrait.

Il est en général schématisé comme suit :



Pipeline du Pentium II

5) Le nombre d'unités d'exécution.

Ils sont tous les trois superscalaires :

Cyrix 6x86 a trois unités d'exécution;

K6 a sept unités d'exécution;

Pentium a cinq unités d'exécution.

Conclusion

Le lecteur pourra utilement se reporter aux [sites des constructeurs](#) pour obtenir des détails supplémentaires.

La conclusion principale est la vérification des conséquences de la théorie fondatrice : quelle que soit la machine, si elle est universelle, elle peut tout faire, et ici en particulier, exécuter le jeu d'instructions x86.

La diversité des architectures concrètes utilisées fait que l'on ne peut plus parler d'architecture induite par le jeu d'instructions. On est en présence d'émulations de l'architecture vue par le programmeur.

2. Processeurs RISC

Le processeur Risc PowerPC 601 est présenté en détail dans l'annexe [asi0003](#). Le tableau suivant donne quelques informations sur plusieurs processeurs RISC. On trouvera ci-après une description succincte du processeur MPC750, annoncé en 1997 par Motorola comme haut de gamme de la famille PowerPC.

Nom	Architecture nbre de bits	Fréquence d'horloge	Cache(s)	Tension d'alim.	Puissance dissipée W.
Alpha21066A	64	166	16k + 0k	3,3	21/50
Alpha21264	64	466 puis 600	112k+0k	2,3	109 à 600MHz
AMD29030	32	25,3	0	5	
AMD9240	32	20,2	4k+2k	5	
PPC602	32	66	4k+4k	3,3	
PPC603e	32	100	16k+16k	3,3	
PPC620	64	133	32k+32k	3,3	30
MPC750P	64	300	32k+32k	1,9	7,9 (max)
SuperSparcII	32	90	20k+16k	5	17
UltraSparc	64	167	16k+16k	3,3	30

3. Processeur Motorola MPC750

Le processeur MPC750, comme son prédécesseur immédiat le MPC740 dont il diffère essentiellement par la gestion du cache de deuxième niveau, est une réalisation de l'architecture Powerpc. Il est de type CMOS, compatible TTL.

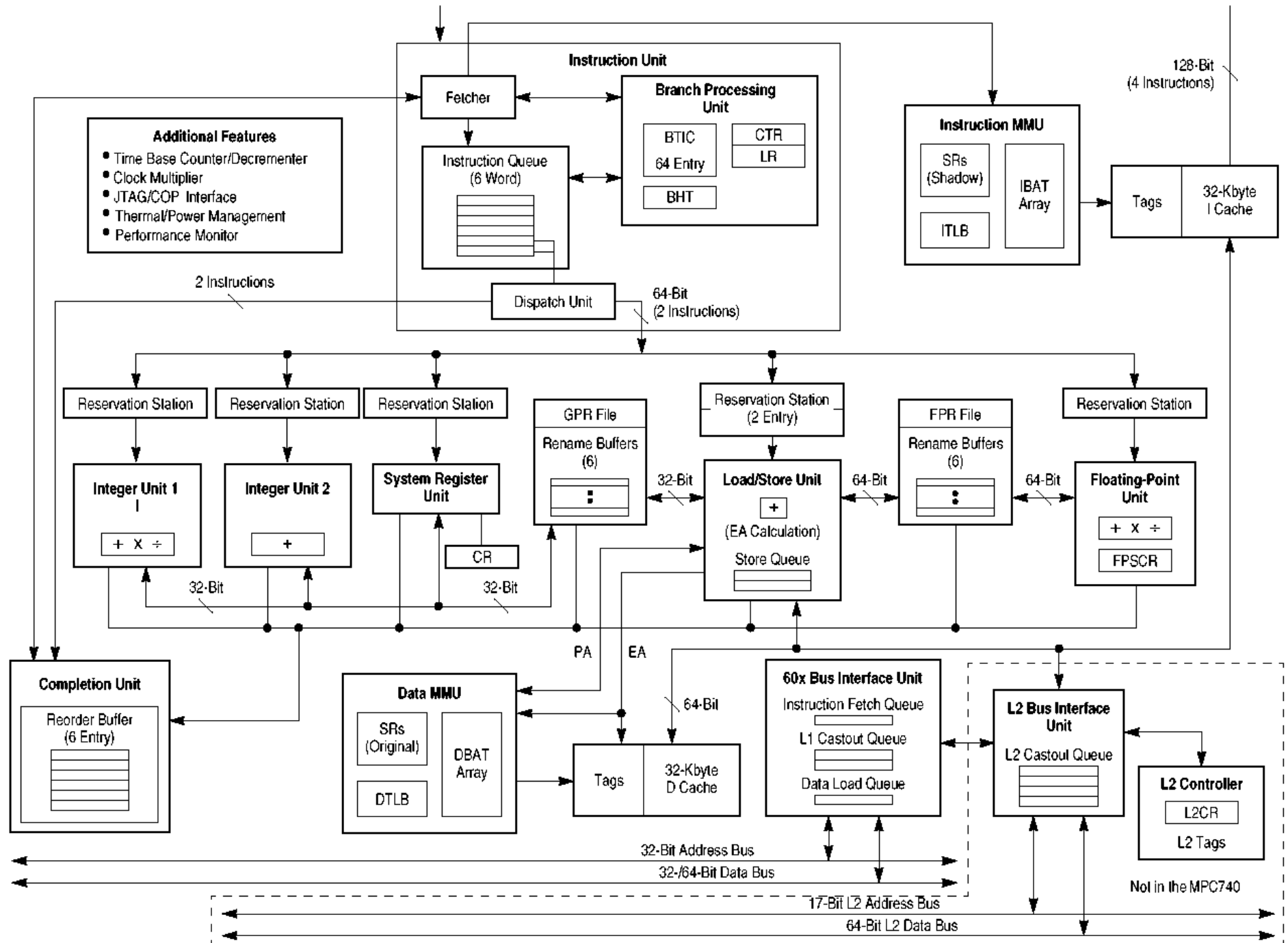
Le cache interne (L1) est commun aux instructions et aux données, 32 Ko, associatif à 8 blocs. Il y a deux UGM pour les instructions et les données. Chaque UGM a 128 entrées, deux tampons associatifs à deux voies (STLB et ITLB) pour sauvegarder les translations d'adresses de pages les plus récentes.

Block address translation is done through the four-entry instruction and data block address translation (IBAT and DBAT) arrays, defined by the PowerPC architecture. During block translation, effective addresses are compared simultaneously with all four BAT entries. The L2 cache is implemented with an on-chip, two-way, set-associative tag memory, and with external, synchronous SRAMs for data storage. The external SRAMs are accessed through a dedicated L2 cache port that supports a single bank of up to 1 Mbyte of synchronous SRAMs.

Il n'y a pas d'interface de cache L2 dans le MPC 740.

Le schéma ci-dessous est repris de Motorola, Advance Information, **MPC750 RISC Microprocessor, Technical Summary, 1997**.

SCHEMA GENERAL DES MPC740 ET MPC750



Annexe 9, Comparaison de processeurs

flottants ont 32 ou 64 bits.

C'est un processeur superscalaire qui exécute deux instructions simultanément.

Il a six unités d'exécution :

- virgule flottante (FPU);
- branchement (BPU);
- système de registres (SRU);
- charge/écrit (LSU);
- deux unités entières, IU1 pour toutes les instructions entières et IU2 pour les instructions entières sauf la multiplication et la division.

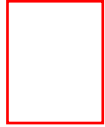
L'unité flottante est en pipeline à trois étages, chacun peut contenir une instruction en simple précision, les instructions en double précision occupent plus de temps.

The instruction unit fetches, dispatches, and predicts branch instructions. Note that this is a conceptual model that shows basic features rather than attempting to show how features are implemented physically. The MPC750 has a 32-bit address bus and a 64-bit data bus. Multiple devices compete for system resources through a central external arbiter. The MPC750's three-state cache-coherency protocol (MEI) supports the exclusive, modified, and invalid states, a compatible subset of the MESI (modified/exclusive/shared/invalid) four-state protocol, and it operates coherently in systems with four-state caches. The MPC750 supports single-beat and burst data transfers for memory accesses and memory-mapped I/O operations. The MPC750 has four software-controllable power-saving modes. Three static modes, doze, nap, and sleep, progressively reduce power dissipation. When functional units are idle, a dynamic power management mode causes those units to enter a low-power mode automatically without affecting operational performance, software execution, or external hardware. The MPC750 also provides a thermal assist unit (TAU) and a way to reduce the instruction fetch rate for limiting power dissipation.



Conservatoire national des arts et métiers
Architectures des systèmes informatiques
ANNEXE 09
Comparaison de processeurs de 1998
Année 2002-2003

Conservatoire National des Arts et Métiers
Architectures des systèmes informatiques
ANNEXE 10
Les choix publics en matière de techniques nouvelles
Année 2000-2001



Le monopole frappe d'immobilisme tout ce qu'il touche.

*Frédéric
Bastiat, Baccalauréat et
socialisme (1848).*

Avertissement au lecteur

Aujourd'hui et plus encore demain, un informaticien doit avoir des lumières sur les choix politiques ou de société en son domaine comme il en a sur les techniques nouvelles. Ces lumières sont utiles à l'homme cultivé, sont nécessaires au professionnel et sont indispensables au citoyen.

L'informatique a été introduite sans grands heurts dans les activités productives jusqu'au milieu des années 1970 bien que les mini-ordinateurs aient modifié certaines façons de procéder dans les années 1960. À l'opposé, la micro informatique a suscité un choix de société. Il n'est que de rappeler les oppositions entre informatique centralisée et informatique répartie, le rapport Nora-Minc sur la téléinformatique centralisatrice et les ouvrages de Bruno Lussato professeur au CNAM, sur l'informatique personnelle dans les années 1970. L'État se mêle d'informatique depuis longtemps.

Les habitudes et les présupposés idéologiques imprègnent les circuits de décision étatiques de notre pays. Ils ont conduit à des échecs techniques et à des désastres financiers. Le texte ci-dessous, ô combien officiel puisqu'il est extrait d'un rapport au Sénat de la République, présente sous une forme concise les plus grandes erreurs commises. Nous le reprenons pour éclairer le lecteur sur les conséquences de ces erreurs qui ont une régularité remarquable et ainsi le mettre en garde pour le futur en procédant comme par vaccination. Le fait que notre enseignement soit donné dans un Conservatoire plaide encore plus pour que la mémoire soit conservée et que le passé serve au futur.

L'auteur aurait pu écrire de son chef un certain nombre de réflexions. Il était bien préférable d'utiliser l'existant et notamment ce texte public. La mine d'informations que constitue les rapports des assemblées est trop peu exploitée. Nous nous essayons à cette exploitation comme on va le lire.

Le document

Le texte ci-dessous est extrait de :

Des pyramides du pouvoir aux réseaux de savoirs, Tome 1,

de Monsieur le Sénateur René Trégouët,

Rapport d'information 331 - 1997 / 1998 - Commission des Finances du Sénat.

Ce rapport est disponible sur le site du Sénat :

http://www.senat.fr/rap/r97-331-t1/r97-331-t1_mono.html

Nous en avons extrait les parties relatives aux choix étatiques en matière de technologies nouvelles, et dans ces parties nous sommes limités à l'informatique.

Pour aider la lecture, les extraits du rapport sont en corps 12, les mentions qui ne font pas partie du rapport sont entre parenthèses, en italiques et en corps 10. Elles ont été ajoutées par l'auteur pour apporter une précision, un commentaire ou une expérience personnelle.

I. L'ACCUMULATION DE MAUVAIS CHOIX EN MATIÈRE DE TECHNOLOGIES NOUVELLES

Sur le plan technique, la société de l'information se caractérise par la convergence des activités de l'informatique, de l'audiovisuel et des télécommunications, grâce, non seulement, à la généralisation du numérique, elle-même permise par la compression des données, mais aussi grâce aux progrès réalisés en matière de composants électroniques (depuis, notamment, l'invention du microprocesseur à l'origine de la micro-informatique).

Or, dans chacun de ces trois domaines, de mauvais choix ont été effectués et des actions mal conduites, particulièrement, hélas, dans le secteur qui, désormais, est le plus important, celui de l'informatique.

A) LES EXEMPLES DE L'INFORMATIQUE

Nos gouvernants et nos élites administratives affectionnent les «plans» ou les «grands programmes», censés mobiliser les énergies et les ressources au profit de grandes causes nationales.

L'élaboration des objectifs de ces actions, pourtant énoncée de façon péremptoire, n'a souvent malheureusement pas été précédée d'une analyse stratégique suffisamment approfondie, qui intègre, notamment, toutes les données du marché.

Trop souvent aussi, leur exécution financière relève d'une logique d'arsenal (l'État est invité à régler les coûts constatés) (*ce type de fonctionnement existe aussi dans l'industrie privée qui travaille pour l'État sous le nom de travail en régie. Un gendarme présent note le nombre d'heures travaillées, le paiement est fait ensuite sur la base d'un coût horaire prédéfini. Le travail de l'arsenal, destiné à garantir la survie de la nation est censé n'avoir pas de prix, tout comme un homme dépensera sa fortune pour assurer sa propre survie : «Mon royaume pour un cheval».*) et le souci de ménager l'argent du contribuable paraît devoir s'effacer devant le caractère prioritaire de la dépense, lui-même justifié par l'importance de l'enjeu.

1. Le Plan Calcul

(Le colbertisme des années gaullistes)

Après le passage de Bull sous contrôle américain à la fin de 1963, (*la Compagnie des machines Bull avait lancé en 1959 une machine très grande machine très novatrice, le Gamma 60. Elle n'a pas eu le succès attendu, en particulier par manque de logiciel, les pertes ont conduit ses propriétaires, principalement le groupe Caillies, à vendre la société à General Electric*), plusieurs administrations s'attachent à définir une politique nationale de l'informatique. (*des tentatives avaient été faites pour «sauver» Bull courant 1963, tout comme dans les années 1930 quand sa vente à IBM avait été envisagée, notamment deux plans avortés, le plan «Quatre axes» qui voulait rapprocher SEA, Bull, CERC I et accessoirement la CSF sous la houlette du ministère de l'industrie puis le plan Hexagone qui voulait regrouper Bull, CSF et CGE sous la houlette du ministère de la recherche*).

Il faut reconnaître que leur tâche est difficile, étant donné l'écrasante suprématie des États-Unis (forts de la domination presque absolue d'IBM et d'un parc informatique représentant les trois quarts de celui du monde).

(Alors que les services de l'armement lancent courant 1964 une étude de définition

d'une gamme de machines, la réflexion passe au CCRST (comité consultatif de la recherche scientifique et technique) dans ce que l'on nomme alors le Comité des Sages, sous la direction de Pierre Lelong, mathématicien spécialiste d'analyse complexe et de géométrie, conseiller de la présidence de la république, plus tard académicien). Il en sortira le rapport Ortolini (alors commissaire général au Plan) dont les recommandations servent de base au Plan Calcul de 1966.

Ce projet, à la préparation duquel aucun représentant du secteur privé n'a été associé, implique la création d'une nouvelle entreprise (la CII : Compagnie Internationale pour l'Informatique) et d'un institut public de recherche (l'IRIA: Institut de Recherche en Informatique et en Automatique).

Le prétexte en est stratégique (*on verra plus tard que ce n'était qu'un prétexte*), les États-Unis ayant refusé de livrer à la France le grand ordinateur scientifique dont elle avait besoin pour concevoir une bombe H.

Mais en fait, cette machine, dont le Plan Calcul n'avait jamais vraiment envisagé la fabrication (*et que la CII ne fabriquera jamais*), nous sera finalement vendue par les Américains.

La Délégation à l'Informatique, nouvelle structure interministérielle directement rattachée au Premier Ministre, chargée de la coordination de l'effort public, ne comprend aucun informaticien professionnel. (*Elle est pour l'essentiel constituée d'ingénieurs fonctionnaires de plusieurs corps, dont un bon nombre d'ingénieurs de l'armement*).

Elle va se heurter, en outre, à l'hostilité des ministères concernés se sentant dessaisis du dossier (Finances, PTT, Industrie, Recherche, Armées...). (*Ceci est très fortement ressenti au ministère de l'industrie et au CNET, beaucoup moins dans les armées*).

«L'échec de la CII est inscrit d'emblée dans ses gènes», estime l'ancien directeur de Bull, Jean-Pierre Brulé.

En effet, les fonctionnaires qui ont conçu le plan pensent en terme d'indépendance nationale plutôt que de viabilité économique. (*Ils utilisent les résultats du contrat de l'armement cité plus haut qui a défini une gamme de machines nommées P0, P1, P2, P3, P4, la série P*).

On rassemble les modestes actifs en informatique des héros malheureux

d'une solution Bull «à la française» (Compagnie Générale d'Électricité, Thomson et Schneider) (*par leurs filiales : SEA concepteur fabricant d'ordinateurs, CAE fabricant d'ordinateurs sous licences et SPERAC fabricant de périphériques*) plutôt que de faire appel au principal constructeur national, en tentant de redresser l'attelage qu'il forme avec General Electric. (*Un autre participant potentiel, très bon fabricant de machines mécanographiques, machines comptables notamment, est Logabax, dont l'actionnaire majoritaire est le groupe belge Electrobél. Cette entreprise n'est pas intégrée dans le plan-calcul pour des motifs d'indépendance nationale. On la retrouvera plus tard dans l'affaire Olivetti.*)

Or, les entreprises de cette troïka (CGE, Thomson et Schneider, maisons mères des précédentes), comptant sur «la traite de la vache à lait publique», pour ne pas passer sous la coupe de constructeurs d'ordinateurs dont la croissance est beaucoup plus rapide que la leur, vont se montrer particulièrement démotivées en tant qu'actionnaires.

Il faut dire que «les ambiguïtés abondent autour du Plan Calcul» - pour reprendre l'expression de J.-P. Brulé - et que la stratégie finalement retenue s'avérera désastreuse.

Elle est en effet marquée par une ambiguïté fondamentale quant au montant des fonds publics engagés (supérieurs au double des 400 MF initialement estimés), de la rentabilité attendue, des produits à fabriquer.

D'un point de vue stratégique, l'erreur consiste à attaquer IBM de front, dans le haut et milieu de gamme des machines de gestion, là où précisément les coûts de fabrication du géant américain sont proportionnellement les plus bas, ses marges les plus élevées et par conséquent ses capacités de résistance à la concurrence les plus fortes. Des constructeurs comme DEC et beaucoup d'autres l'avaient bien compris (*de même les entreprises dont on force la fusion ; CAE fabriquait des ordinateurs scientifiques sous licence américaine de TRW puis de Scientific Data Systems, SEA concevait et fabriquait les machines de bureau CAB et Sperac fabriquait des périphériques*).

Certes, les besoins des administrations nationales sont importants dans ce secteur des grosses machines de gestion (*la CII recevra un quasi monopole de fournisseur des administrations, enseignement supérieur inclus, contraintes de lui acheter leurs ordinateurs*) mais, d'une part, Bull-GE y est aussi présent, et d'autre part, il ne s'agit que de 1% du marché mondial (dont le marché français représente 5%). (*Ces années sont celles*

de l'ouverture du marché des miniordinateurs que fabrique d'abord Digital Equipment Corporation).

La CII réussit cependant, contre toute attente, son démarrage technique dans la période 1967-1971 (elle correspond à sa première convention avec l'État) et la commercialisation de ses produits Iris (seulement partiellement compatibles entre eux!) est correctement engagée. (À ses débuts, et avant de produire les machines de la série P, la CII abandonne les CAB originales et poursuit la fabrication commencée par CAE de ses Sigma 2 et Sigma 7, sous licence de l'américain Scientific Data Systems sous les noms de 10020 et 10070. Elle produit le 10020 en novembre 1967. L'IRIA, dans le même temps se lance aussi dans la conception d'au moins un ordinateur, en l'espèce la machine Q0 nommée ensuite Myria qui n'a pas eu suite immédiate. La CII lance la conception de la gamme Iris, nom commercial de la série P. L'Iris 50 est produit courant 1968, l'IRIS 80 en septembre 1969. Une autre machine très réussie de la CII fut le Mitra, mini ordinateur microprogrammable, produit seulement à partir de 1972 dans plusieurs versions civiles et militaires.

La CII a fabriqué le 10010, comparable au Myria, totalement incompatible avec les 10020 et 10070, malgré ce que laissait croire leurs noms. La conséquence fut particulièrement sensible à la RATP. Celle-ci avait entrepris de gérer par ordinateurs les passages des voyageurs munis de tickets magnétiques sur la ligne A du RER, la première construite. Pour cela un 10020 avait été acheté, avec un 10010 en secours. Il fallut refaire la totalité de la programmation, alors en assembleur, pour la machine de secours, cela donna lieu à un mémoire d'ingénieur du CNAM).

Sa croissance est forte (supérieure à 25 % par an) et ses effectifs (6.000 personnes) atteignent la moitié de ceux de Bull en France en 1971.

Toutefois, les coûts de fabrication et de distribution de ses produits excèdent leurs prix de vente. La CII vend ainsi à perte bien qu'on lui ait entièrement payé les études de ses machines.

La première convention avait coûté à l'État environ 1 milliard de francs, la deuxième (1972-1975) sera dotée de plus du double et reviendra en fait à l'État à trois fois ce montant, tandis que l'apport des actionnaires sera, quant à lui, 2,5 fois moins élevé qu'en 1967.

C'est alors (en 1972) qu'est envisagée une sortie européenne à ce plan mal engagé.

Il faut le reconnaître : l'idée de départ n'est pas mauvaise en soi :

associer CII, doté d'un bon bagage technique et fort de sa ligne de produits Iris, au géant allemand Siemens, licencié du groupe américain RCA qui vient de jeter l'éponge en ce qui concerne la fabrication d'ordinateurs.

On pense par ce biais créer une synergie entre la puissance commerciale et financière de Siemens et la technique de CII, dont on espère qu'elle se verra ouvrir le marché européen, et en particulier allemand.

Il ne semble pas pour autant qu'ait été envisagé aucun scénario alternatif susceptible, notamment, de donner accès au marché américain, beaucoup plus substantiel : autant dire qu'on retrouve à l'œuvre toujours la même insuffisance de réflexion stratégique.

La CII, qui ne fait que de l'informatique, ne pèse pas lourd face à un groupe aussi fort et diversifié que Siemens. Ses intérêts vont en outre être mal défendus. Les négociations seront menées de façon dispersée, tantôt par la compagnie elle-même, pour les projets opérationnels de coopération, tantôt, en ce qui concerne les intérêts des actionnaires, par Thomson, paralysé par les réticences de la CGE et désireux avant tout de faire payer par l'État la nouvelle addition qui se profile.

La CII cède aux exigences de Siemens, relatives à la compatibilité des futurs produits communs aux deux groupes avec ceux d'IBM (ce qui risque de l'obliger à concevoir des machines selon des schémas et peut-être des brevets appartenant à son associé) et se voit proposer, au prix fort, une participation au réseau commercial du groupe bavarois en dehors d'Allemagne.

La Délégation à l'Informatique va jusqu'à envisager alors de faire financer par le contribuable français cet achat de filiales étrangères du grand groupe allemand.

Mais la note sera finalement réduite et partagée avec l'État allemand, non sans que la France ait exigé une structure d'organisation assez souple pour que d'autres constructeurs puissent par la suite s'y associer.

Philips en profite pour rejoindre le consortium, dénommé Unidata, en tant que partenaire à part entière avec un apport pourtant des plus limités.

Selon Jean-Pierre Brulé, la formule d'association adoptée se révèle «cauchemardesque».

Le modèle d'organisation retenu, lourd et mal coordonné, est incompatible avec les exigences imposées par la conception et la réalisation d'une ligne de produits compatibles entre les trois partenaires.

«Unidata cumulait les handicaps - écrit l'ancien président de Bull - les passés techniques divergents, les différences d'appréciation quant à l'urgence d'une nouvelle ligne, la complexité de la structure de prise de décisions, l'absence de connaissances du sujet au niveau où les arbitrages remontaient, enfin, des conflits d'intérêts permanents». *(Le même aurait pu tenir les mêmes propos au sujet de la fusion, jamais vraiment réussie de SEA et CAE où les anciens de CAE, fabricants sous licence eurent le pas sur ceux de SEA, concepteurs de machines.)*

De sorte que, très vite, Siemens et Philips vont s'apercevoir qu'une véritable fusion est la condition du succès.

Mais, du point de vue des intérêts français, une telle solution revient à faire financer, une fois de plus, par de l'argent public, une participation au nouvel ensemble intégré, vouée à demeurer, en tout état de cause, minoritaire.

C'est en effet la débandade du côté des actionnaires de la CII : Schneider cherche à vendre ses parts, la CGE a refusé l'augmentation de capital qui devait accompagner l'entrée dans Unidata. Elle désavoue donc les accords négociés par Thomson.

La cacophonie française atteint son comble lorsque les pouvoirs publics, sans avoir mis au courant la Délégation à l'informatique, *(pendant toutes ces années cohabitaient la Délégation à l'informatique et la très influente DIÉLI, direction des industries électriques, électroniques et informatique du Ministère de l'industrie. Une nouvelle structure étatique, l'Agence pour le développement de l'informatique, sera créée en 1982 et dissoute en 1986)* commencent à songer à un rapprochement entre la CII, qui continue à négocier avec Siemens et Philips, et Bull, désormais associé à l'américain Honeywell. *(après que General electric se soit retiré de l'informatique.)*

Un démenti est opposé aux partenaires européens de la compagnie au sein

d'Unidata qui s'inquiètent des rumeurs de contacts pris dans ce sens.

Ceux-ci auront l'impression d'avoir été floués, lors de l'annonce, en mai 1975, des accords (*de fusion*) CII-HB qui sonnent le glas des espoirs de construction d'une industrie européenne de l'informatique. (*Il n'y a plus alors qu'un seul fabricant généraliste en France, nommé CII-Honeywell-Bull*)

«Pour le contribuable, la note des deux conventions du Plan Calcul et de leur sortie se monte à environ 13 milliards de francs 1992». Certes, ajoute J.P. Brulé, «l'apport de la CII permettra, en 1976, de ramener Bull sous majorité française. Mais la même chose était probablement faisable en 1966, et pour 10 à 100 fois moins d'argent».

Bref, il en a coûté moins que l'équivalent d'une journée de coût du Plan Calcul pour lancer, à partir de fonds strictement privés, des constructeurs comme DEC ou Apple, deux des plus grands succès de l'informatique mondiale, des années 70 et 80.

«L'informatique malade de l'État» : tel est le titre de l'ouvrage, précité, de Jean-Pierre Brulé. Malheureusement la liste des exemples qu'il cite, à l'appui de ce diagnostic, ne s'arrête pas à la fusion de la CII et de Honeywell-Bull.

(*À titre de compensation, la Télémécanique est vendue à Thomson, c'est la fin de l'informatique dans cette entreprise*).

2. Le raid manqué sur Olivetti

(Les incohérences prétendument libérales)

Saint-Gobain, groupe spécialisé dans le verre et les tuyaux de fonte, fait sensation en annonçant en 1978 sa participation au «Plan composants», nouveau grand programme à la française (*Le plan composants est destiné à maintenir et développer une industrie des composants électroniques, ce que le plan calcul avait oublié. Il s'agit essentiellement d'une création de la DIÉLI déjà citée*). Il confirme, un an plus tard, ses nouvelles ambitions dans le domaine de l'électronique au sens large (y compris l'informatique) en rachetant la participation de CGE dans la Compagnie des Machines Bull, dont il devient le principal actionnaire avec l'État.

Cette opération ouvrant à Honeywell un droit automatique de sortie du

capital de Bull (obtenu en 1975, par la CGE, pour lui donner une influence sans commune mesure avec sa participation au capital de notre constructeur informatique national), un fonds pour le rachat éventuel des actions de la firme de Minneapolis est constitué en partie avec de l'argent public, en partie avec des fonds du verrier français (on parle alors du «milliard de Saint-Gobain »).

Une série de concessions sont consenties au groupe américain pour qu'il accepte le remplacement de CGE par Saint-Gobain (Bull est prié, notamment, de renoncer à ses griefs concernant les entorses de Honeywell au principe d'une ligne commune de produits).

L'année suivante, en 1980, Saint-Gobain annonce l'acquisition du tiers environ du capital d'Olivetti, qui va lui permettre, ensuite, de prendre le contrôle de Bull grâce à une opération boursière montée avec l'accord de l'État (des actions Olivetti sont échangées contre des actions CMB (*l'ancienne compagnie des machines Bull*) possédées par de petits porteurs).

Là encore, il faut le reconnaître : comme celle du rapprochement avorté entre CII et Siemens, l'idée d'une coopération entre Bull et Olivetti est sensée : acteur incontournable de la bureautique mondiale, Olivetti est handicapé par ses faiblesses dans le domaine des logiciels et de l'architecture de systèmes. Bull peut l'aider à les surmonter; et en échange, le constructeur italien peut offrir au français des débouchés en bas de gamme.

Une fois encore, ce sont les considérations stratégiques et la conduite du déroulement pratique des opérations qui vont se révéler pour le moins déficientes.

Le montage effectué va en effet avantager Olivetti aux dépens de Bull, en allant jusqu'à lui sacrifier Logabax, pionnier français de l'ordinateur de bureau depuis 1967. (*Cette entreprise déjà citée, dont l'actionnaire majoritaire est le groupe belge Électrobel a toujours eu une grande indépendance, elle avait créé les premières imprimantes à aiguilles, les LX 180 et 200 et fabriquait les micro ordinateurs LX 500, 515 et 525*).

L'essentiel du milliard de Saint-Gobain, quant à lui, va être affecté à la prise de participation d'un tiers dans le capital d'Olivetti.

Déjà affaibli par les concessions faites à Honeywell (cf. supra), Bull,

sous-alimenté depuis 1976 par ses actionnaires, continue de manquer de capitaux propres.

Son passage sous contrôle de Saint-Gobain, qui n'a pas coûté un centime à ce dernier, ne lui rapportera pas un sou d'argent frais.

Ayant dû renoncer à ses projets d'acquisition en cours (Olympia-Hermès), Bull est écarté de la reprise de Logabax (nous y reviendrons).

Il ne bénéficiera d'aucun contrat d'études public pour la bureautique au titre du CODIS (Comité de développement des industries stratégiques) (*nouvelle structure étatique de distribution de l'argent public*) jusqu'à la fin de 1981.

Les dirigeants de Saint-Gobain, et avec eux les représentants de l'administration française vont, en fait, se laisser magistralement berner par Carlo de Benedetti.

Nullement désireux de collaborer avec Bull, Olivetti poursuit en réalité deux objectifs : le renforcement de ses fonds propres et l'accès à des contrats d'études publics français.

Saint-Gobain va combler son attente sur le premier point. Il contribue, en effet, à désendetter Olivetti en souscrivant à une augmentation de son capital.

La déception rencontrée, en revanche, par la firme d'Ivrea, en ce qui concerne la réalisation du deuxième point, va conduire Carlo de Benedetti à la rupture, conduisant ce dernier à multiplier les déclarations provocantes du genre :

«Que pourrait bien apporter à Olivetti un groupe spécialisé dans le verre et les tuyaux de fonte?»

Ou bien : «Saint-Gobain détient 33 % de notre capital, mais je conduis mon entreprise de façon totalement indépendante».

Cela n'empêche pas le ministère de l'industrie de donner son accord à la reprise par Olivetti, de préférence à Bull, de la société française Logabax, spécialisée pourtant non pas dans la bureautique, mais dans l'informatique de gestion.

Vidée de sa substance, c'est-à-dire de ses capacités propres d'études, Logabax finira comme un simple département commercial d'Olivetti. (*La société R2E pionnière de la microinformatique avec ses machines Micral, a été de la même façon achetée et totalement stérilisée par Bull en 1982*).

Malgré sa complaisance envers elle, Saint-Gobain n'exerce en effet pratiquement aucune influence sur la gestion de l'entreprise italienne - comme Carlo de Benedetti l'a fait remarquer publiquement- en raison de la fragmentation de cette dernière et du contrôle de 40% de son capital par un syndicat à l'italienne.

L'administration française n'ira jamais vérifier la réalité de l'option sur les actions personnelles de Carlo de Benedetti que Saint-Gobain prétend détenir et qui serait susceptible de lui donner le contrôle majoritaire d'Olivetti. Et seul Honeywell semble s'inquiéter du danger d'un éventuel transfert de savoir-faire en logiciel de Bull vers Olivetti, sans contrepartie.

Tout cela n'a pourtant rien de surprenant dans la mesure où -comme le note J.P. Brulé- le schéma retenu a placé Saint-Gobain, qui ne possède ni expérience du métier, ni dirigeants formés à une tâche de cette ampleur, en coordinateur obligé de deux très importants investissements en informatique : Bull et Olivetti.

«Les pouvoirs publics, si souvent tatillons, font preuve en l'occurrence -poursuit l'ancien président de Bull- d'une énorme confiance dans le groupe verrier promu au rang de champion national des puces électroniques et de l'ordinateur».

Quelques mois plus tard, (*nationalisations socialistes de 1982*) le Gouvernement sépare Bull et Saint-Gobain, tous les deux nationalisés, et contraint ce dernier à revendre ses parts d'Olivetti.

La démonstration de l'incohérence et de l'incompétence des pouvoirs publics semble avérée, mais le contribuable n'en est pas encore au bout de ses peines !

3. La nationalisation mal conduite de Bull

(Les joies du mécano industriel ou le volontarisme socialiste)

La nationalisation de Bull, en avril 1982, s'effectue dans un contexte de détérioration de ses résultats.

Cette dégradation est due, selon son ancien Président, à la concurrence d'IBM et à la politique de Saint-Gobain.

Concurrencé par les premiers mini-ordinateurs, le géant américain mène une politique de guerre des prix, en milieu de gamme.

Bull riposte par une politique tarifaire agressive dans la commercialisation de l'excellent mini-ordinateur «Mini 6» conçu par Honeywell, mais les capacités de l'usine d'Angers ne permettent pas de faire face à la croissance des commandes.

La réduction des marges rend plus difficile l'autofinancement des investissements nécessaires, au moment même où les apports en capital des actionnaires décroissent, jusqu'à devenir nuls en 1980, année du remplacement de la CGE par Saint-Gobain. En outre, ce dernier exige une augmentation substantielle des dividendes versés par le constructeur informatique dont le total des emprunts, dans ces conditions, dépasse, à la fin de 1980, celui des fonds propres.

Du printemps 1981 à celui de 1982, Bull, en attente de nationalisation, traverse une année noire, contraint d'embaucher (*politique socialiste de l'emploi non productif*), privé d'apport en capitaux par Saint-Gobain, toujours confronté à des difficultés techniques dans la fabrication de son nouvel ordinateur DPS 7, tandis que sa clientèle s'inquiète de l'avenir de son partenariat avec Honeywell.

Les commandes chutent à partir de juillet, et la compagnie connaît en 1981 sa première perte depuis dix ans.

La nationalisation de Bull entraîne une négociation serrée avec Honeywell qui en détenait 47 % et qui accepte de n'en conserver que 20 %. Il en résulte pour le contribuable français une charge de 150 millions de dollars et, pour la compagnie, une dépense annuelle de 37 millions de dollars (licences + dividendes).

Pour l'ancien Président de Bull, démis de ses fonctions en juillet 1981, plusieurs erreurs graves sont imputables au nouveau pouvoir : il laisse Bull dériver et encaisser les mauvais coups de son actionnaire

Saint-Gobain, le pousse à embaucher, malgré ses pertes, ne sépare le destin des deux groupes qu'après leur nationalisation, et négocie, enfin, avec Honeywell dans de mauvaises conditions initiales.

Avant la nationalisation, le bilan global de la compagnie est honorable mais elle n'a jamais trouvé un actionnariat qui la soutienne, sans arrière-pensées, dans sa croissance, et lui apporte les fonds propres dont elle avait besoin.

D'autre part, elle n'a pas suffisamment investi et surtout n'a pas remis en cause sa stratégie d'imitation du modèle IBM alors que la question d'une réorientation vers les mini et micro-ordinateurs et vers la bureautique méritait, au moins, d'être posée (*ne fût-ce qu'à l'occasion de l'achat de R2E*).

Dans les dix années suivant sa nationalisation (de 1982 à 1992), Bull a reçu l'équivalent de 20 milliards de fonds publics (dotation en capital + aides aux études). Cela ne l'a malheureusement pas empêché ni de procéder à des milliers de licenciements, ni de voir diminuer son niveau de maîtrise technologique.

En outre, alors que la croissance des ventes mondiales d'ordinateurs personnels a été très soutenue pendant plusieurs années consécutives, Bull semble avoir «misé sur le mauvais cheval» en acquérant l'américain Zenith Data Systems (cédé par la suite à Packard Bell) dont les pertes continuelles ont lourdement pesé sur ses résultats.

4. Le mirage de la filière électronique

(ou la poursuite du volontarisme)

Publié au printemps 1982, le «rapport Farnoux» (*Abel Farnoux est un très proche conseiller d'Édith Cresson, amie proche du président de la république, alors ministre de l'agriculture et future premier ministre*) est à l'origine d'un programme très ambitieux baptisé PAFE (Programme d'action pour la filière électronique).

Parmi les onze branches concernées figure l'informatique pour laquelle on recommande tout à la fois la poursuite de l'effort concernant les systèmes moyens, le lancement d'un grand ordinateur scientifique, la production de masse de micro-ordinateurs et le développement de «briques de base» (modules standardisés utilisant notamment les mêmes

puces) pour la mini-informatique.

Les dépenses correspondantes sont estimées à 140 milliards de francs sur cinq ans, somme faramineuse, mais on s'aperçoit qu'il s'agit d'un amalgame, budgétairement hétérodoxe, de dépenses hétérogènes (fonctionnement et investissement, mesures nouvelles et acquises; budgets civils et militaires, publics et privés, budget annexe des PTT, etc. (*ce budget annexe des PTT était le faux nez qui couvrait les bénéfices colossaux de la partie téléphonique des PTT qui, fort de son monopole, plumait consciencieusement la volaille, particuliers et entreprises.*))

Sept projets nationaux censés provoquer un saut technologique et associer étroitement les industriels aux laboratoires publics sont lancés immédiatement. Mais il s'agit, en fait, d'un cocktail composite, allant de la traduction assistée par ordinateur aux «briques de base» pour mini machines mentionnées ci-dessus.

Finalement, les résultats de ce vaste programme, qui ambitionnait de faire de la France «la troisième puissance technologique de l'an 2000», juste derrière les États-Unis et le Japon, s'avèrent très décevants. Presque tous les objectifs solennellement affichés n'ont pu être atteints. L'accroissement des budgets de recherche concernés est, au total, très limité.

Dans le budget de l'Industrie, l'électronique continue de passer après les houillères nationales, en ce qui concerne les subventions, et de venir après les chantiers navals, la sidérurgie et Renault, s'agissant des dotations en capital.

À l'examen, deux erreurs d'appréciation majeures expliquent ces déconvenues.

La première est illustrée par l'absorption par Bull, à l'automne 1982, des filiales informatiques de Thomson et de la CGE. Elle consiste à tout miser sur un «champion national», fabriqué artificiellement par une sorte de mécano industriel dirigiste, à l'inverse de la politique suivie alors avec succès par le Japon.

La deuxième erreur, quant à elle, consiste à pécher par excès d'orgueil en croyant que nous avons les moyens de maîtriser, à travers le budget de l'État, tous les éléments de la filière électronique dans son ensemble.

Depuis trente ans, fait observer Jean-Pierre Brulé dans «L'Informatique malade de l'État», les plus brillants succès ont été remportés par des sociétés qui se sont spécialisées dans des créneaux, qu'il s'agisse des mini (DEC), des micros (Apple) et de leurs logiciels (Microsoft) ou des machines géantes (Cray). Nous parlerions aujourd'hui de Sun, Compaq et toujours, bien entendu du binôme Wintel (Microsoft et Intel).

À l'inverse, les constructeurs généralistes américains et européens n'ont cessé de perdre du terrain. Seul NEC, parmi ces derniers, se prévaut encore de la triple intégration communications-ordinateurs-composants, même si, chez plusieurs d'entre eux comme IBM, les composants épaulent l'informatique.

Les sociétés françaises qui ont le mieux réussi ont prospéré, sans l'aide de l'État, dans des créneaux qu'elles ont elles-mêmes trouvés, se trouvent essentiellement dans le logiciel (Business Object...), les services (Cap Gemini...) ou les cartes à puce (Gemplus...).(Le lecteur remarquera que l'État n'a jamais établi de plan logiciel ou de plan services même si cette idée a été évoquée plusieurs fois et notamment en 1982 où la création d'une entreprise nationale par fusion de sociétés de SSII était examinée. Elle aurait porté le nom d'Informatique de France. Il eut été à craindre que s'il l'avait fait, ces secteurs eussent été eux aussi en piètre état. Et pourtant, on pouvait raisonnablement craindre ce plan logiciel dont Maurice Allègre, ancien délégué à l'informatique donne les motifs en 1982 :

«Les Français ont-ils le génie logiciel?», «Heureusement, nous avons les logiciels», «Les Français sont faits pour faire du logiciel», «Nous occupons la deuxième place dans le monde du logiciel»... entend-on souvent dire. Ces affirmations sont en partie vraies, car l'industrie française du logiciel est particulièrement dynamique, reconnaît Maurice Allègre, directeur de la Desti, lors de l'allocution d'ouverture du Colloque de génie logiciel, cette semaine à Paris. Cependant, poursuit-il, d'importantes questions restent sans réponse. «Comment se fait-il, par exemple, qu'aussi peu de logiciels français soient commercialisés en quantités importantes? Pourquoi une grande partie de l'activité de vente de logiciels de nos SSII est réalisée avec des logiciels d'origine américaine? Quelles seront les retombées du langage Ada sur le chiffre d'affaires de l'industrie française du service informatique?. Autant de questions qui laissent planer un doute sur notre réelle aptitude à produire du logiciel vendable et exportable...»

Le Monde Informatique du 14 juin 1982)

Malgré (ou peut-être à cause) des actions maladroites de soutien de l'État aux créateurs français de logiciels (point faible de notre pays), la France est aussi absente de ce secteur stratégique puisque

l'industrie des logiciels universels pour micros est aujourd'hui exclusivement américaine.

Parmi les succès rencontrés dans ce domaine, il convient de mentionner celui de Borland, société californienne fondée dans les années 80 par un universitaire français, Philippe Kahn. «S'il avait trouvé en France - s'interroge J.-P. Brulé - au lieu des coquecigrues de la filière électronique, un terrain plus stimulant, plus fertile sur les plans économique et culturel, n'y aurait-il pas fondé son entreprise ?». (*P. Kahn était l'auteur du logiciel de base des premiers Micral de R2E.*)

On ne demande pas à l'État de choisir les créneaux ni les stratégies, conclut-il. «On se satisferait bien qu'il crée les conditions de leur éclosion». En d'autres termes, une aide intelligente de l'État consiste à amender le terrain des créateurs d'entreprises (*au sens de l'amendement des terres agricoles : apport de matières pour modifier leur composition et les rendre plus fertiles, telles que phosphates, limons, etc. Elle n'a aucun rapport avec les pesticides contraignants et avec les manipulations génétiques.*).

5. Le plan informatique pour tous

(cela n'a pas marché, alors continuons !)

Déjà évoqué dans la partie du précédent chapitre consacrée au rôle des NTIC (*nouvelles technologies de l'information et de la communication*) dans l'éducation, le plan informatique pour tous constitue un exemple de plus des gabegies auxquelles conduit un certain volontarisme étatique.

Ce plan visait à mettre en place, dès la rentrée de 1985, dans 50.000 établissements, plus de 120.000 machines et à assurer la formation, pour la même date, de 110.000 enseignants. Son coût était évalué à 1,8 milliard de francs (1,5 milliard pour le matériel, le reste pour la formation et les logiciels). (*La configuration de base était : une machine 16 bits, entourée de machines 8 bits techniquement et fonctionnellement obsolètes, en réseau local. Une simple division montre que la dépense de formation et d'achat de logiciels se montait à 3000 F par enseignant en un temps où le prix du marché de la formation était de l'ordre de 4000 F. par jour*)

Rappelons-le une fois encore : ce ne sont pas les objectifs en cause qui prêtent à la critique mais leurs modalités financières et pratiques de mise en œuvre.

Sur le plan financier, tout d'abord, les crédits budgétaires se sont révélés insuffisants, compte tenu de l'importance des dépenses à engager et des délais impartis. De sorte qu'il a fallu recourir à la formule du crédit bail (*pour les achats de matériels*), incompatible avec le principe de l'annualité budgétaire et faire intervenir, irrégulièrement, le budget annexe des PTT pour payer les annuités correspondantes. (*redisons le : le budget annexe des PTT est l'expression neutre qui désignait alors le bas de laine accumulé année après année grâce aux tarifs prohibitifs du téléphone en France*)

Sur le plan pratique, en raison du caractère fermé et propriétaire de l'Operating System (OS), des retards de plusieurs mois ont été enregistrés dans la livraison des logiciels, entraînant démotivation et pertes de savoir-faire chez les enseignants par ailleurs trop rapidement formés. Ces derniers, dans leur grande majorité, estiment ne pas avoir été mis en mesure d'utiliser valablement l'informatique dans l'accomplissement de leur mission éducative. Ils jugent, en outre, le temps passé devant les consoles disproportionné par rapport à l'acquis pédagogique escompté. (*Deux types de formation ont été envisagées pour les enseignants, une formation lourde de plusieurs mois, qu'il était impossible de donner à 50000 enseignants, ce qui d'ailleurs n'en aurait fait qu'un par établissement, et une formation légère de quelques jours qui s'est avérée insuffisante. Il n'existait évidemment pas de ligne rouge qu'un enseignant appelle en cas de difficultés.*)

Mais la plus grande erreur du plan informatique pour tous fut de s'appuyer sur un «operating system» (OS) propriétaire. Cette démarche opposée à toute l'approche mondiale interdisait d'utiliser tous les logiciels compatibles Apple ou Dos qui étaient alors, déjà, les plus usités. Aussi, les fabricants de contenus ne voulant pas se lancer dans de lourdes productions pour un marché trop étroit, les enseignants ne disposèrent que de quelques maigres produits à présenter à leurs élèves.

Résultat : L'utilisation des ordinateurs n'atteint même pas trois heures par mois et par élève dans plus de 60% des établissements. La qualité et l'utilité des logiciels sont mises en cause avec un pourcentage d'insatisfaction qui atteint 55% dans les lycées.

Les contrats obtenus par l'industrie française ne lui permettront même pas de rattraper tant soit peu son retard en micro-informatique qui est notoire : malgré la commande, dans le cadre de ce plan, de près de 100.000 machines, Thomson (*fournisseur des machines à 8 bits T07 et M05*)

abandonne la partie en 1989 ainsi que Matra. La CGCT sera vendue à Ericsson et Léanord à Siemens. (Les machines T07 et M05, livrées à partir de 1985 étaient basées sur des processeurs à 8 bits très largement dépassés et royalement équipées de lecteurs de cassettes. Il avait été considéré par les décideurs que ces microprocesseurs étaient largement suffisants pour des écoliers et des collégiens. Pour relier ces machines, la société Léanord crée un Nano-réseau, totalement incompatible avec le reste du monde informatique. Les enseignants très mal formés ont alors à maîtriser non seulement des machines mais en plus un réseau.)

Une deuxième vague succède, en 1988, au plan informatique pour tous avec l'opération «13.000 micros pour les collèges et les lycées» (il s'agit enfin de machines à processeurs de 16 bits). Le montant budgétaire prévu est cette fois plus modeste (100 MF) et il est fait appel à du matériel haut de gamme, installé par l'UGAP (Union des groupements d'achats publics, organisme placé sous la tutelle du ministère de l'Éducation nationale). 60% des nouveaux ordinateurs achetés par l'Éducation nationale sont désormais d'origine étrangère.

Les priorités vont à l'équipement des collèges, à l'enseignement technologique (bureautique et productique) ainsi qu'aux classes préparatoires à certaines grandes écoles (biologie, HEC...).

Mais une enquête réalisée en 1992 révèle qu'à peine 15 à 20% des enseignants déclarent utiliser l'informatique en classe.

Une conclusion s'impose : déconcentrer les équipements qui deviennent rapidement obsolètes et en transférer ainsi le coût de renouvellement de l'État aux collectivités locales risque de créer de fortes disparités.

(Le principal problème du plan informatique pour tous fut le manque de formation : «Je me souviens d'un professeur formé qui confondait joyeusement les bits et les bytes...» dit un témoin.

M05 Micro-ordinateur construit par Thomson, sorti en Juin 1984, n'eut de succès que dans le plan informatique pour tous.

T07 Micro-ordinateur construit par Thomson est sorti au Sicob de 1982, bien après les premiers PC. Il était muni d'un clavier repoussant et d'un lecteur de cassettes. Le plus souvent, il est resté dans des cartons après livraison dans les lycées et les collèges.

Le plan informatique pour tous a pourtant laissé de bons souvenirs à des hommes politiques et flattait leur goût pour les structures étatiques, témoin une déclaration de M. Lang :

«S'agissant des nouvelles technologies, je crois qu'il faudrait agir à la fois sur les équipements et sur les contenus. Sur les contenus, je crois que parfois des formules simples, comme par exemple la création d'une agence pour les nouvelles technologies, qui serait animée par un petit commando doté de moyens financiers relativement modestes, mais qui aurait la confiance des responsables, pourrait soutenir l'ensemble des recherches dans tous les domaines, des CD Rom à la création de sites nouveaux, en passant par la numérisation de l'ensemble des bibliothèques de France, etc. Sur les équipements, je pense qu'il faudrait s'inspirer de ce qui a été fait en 1985 à travers le plan Informatique pour tous, qui a été critiqué mais qui a quand même réussi à imprégner ou à immerger les écoles dans la culture informatique. On pourrait imaginer un plan micro-ordinateur pour tous qui consisterait par exemple à confier à chaque professeurs 2 micros, un pour lui, pour qu'il s'y exerce chez lui, et un autre à l'école, pour qu'il initie les élèves ou d'autres collègues.»

6. Le projet Sesam Vitale

Le projet Sesam Vitale d'informatisation de l'assurance maladie constitue le dernier et le plus récent exemple d'opération d'envergure mal conduite.

À l'horizon de l'an 2000, l'ensemble des ressortissants du régime général doivent se trouver dotés d'une carte à puce remplaçant le support papier actuel ainsi que le carnet de santé et comportant, éventuellement, des données relatives à l'assurance complémentaire.

Un système de télétransmission des informations du cabinet du médecin vers les caisses devait se substituer, avant la fin de 1998, aux 850 millions de feuilles de soins circulant chaque année, ce qui suppose la mise en place d'ici cette date des équipements et liaisons informatiques nécessaires. Mais le mécanisme prévu ne couvrirait ni le règlement des honoraires ni les prescriptions (continuant à être établies sous forme d'ordonnances).

Ces échanges permanents d'information entre les caisses, les professionnels de santé et les assurances sont indispensables à une maîtrise efficace des dépenses selon le directeur de la CNAMTS, Gérard Rameix.

Le coût de la première phase de la réalisation de ce projet est évalué à 4 milliards de francs.

Engagé depuis dix ans, sans volonté politique ferme, ce processus a été

relancé en 1996 par le plan Juppé de réforme de la sécurité sociale.

Malgré de tels délais, M. Rozmaryn, dans un rapport remis au Gouvernement à la fin du mois d'octobre 1996, a estimé que cette révolution avait été mal préparée par l'assurance maladie.

L'ingénieur général des télécommunications y dénonce tout à la fois des lacunes dans la sécurité du système «peu arbitrée et sans doute mal équilibrée» et surtout l'absence de retour d'information vers les professionnels de santé pourtant soumis à des contraintes alourdies. Il aurait préféré confier la maîtrise d'ouvrage à un délégué général entouré d'une petite équipe de haut niveau plutôt qu'à la CNAMTS, comme l'a décidé le Gouvernement.

En revanche, satisfaction lui a été donnée en ce qui concerne la création d'un «Conseil supérieur des systèmes d'information de santé» correspondant à la structure de concertation qu'il souhaitait.

M. Rozmaryn suggérait de ne pas faire cette réforme «à marche forcée». Dans ces conditions, il est peu probable que l'informatisation des cabinets médicaux soit achevée à la fin de 1998 comme l'avait annoncé le précédent Gouvernement.

Au moins les deux derniers exemples cités concernent-ils des domaines (éducation et santé) qui sont par nature du ressort de l'intervention des administrations publiques.

Tel n'était pas le cas des ingénieurs de l'État dans la sphère des activités industrielles marchandes.

Plus préoccupés de tendre la sèbile ou de limiter les dégâts causés par les interventions trop souvent maladroites des pouvoirs publics, les constructeurs informatiques français n'ont pas su définir les stratégies qui leur auraient permis de mieux s'adapter aux évolutions de la technologie et du marché.

Cela s'est produit lors de l'irruption de la micro-informatique et de l'essor des progiciels (dont la diffusion ne connaît pas de limites physiques, contrairement aux activités de services qui se comptent en heures d'ingénieurs et dépendent de l'existence de la demande d'une clientèle particulière).

Souhaitons, même s'il est un peu tard, que le maximum de nos PME sachent profiter de l'explosion des réseaux sous toutes ses formes (Internet, Intranet...), sans entrave de la part de l'État (comme nous le verrons plus loin).

En tout état de cause, jusqu'en 1992, les actions de l'État dans le secteur de l'informatique, tous types d'interventions confondus, auront coûté au contribuable français plus de 40 milliards de francs pour un bien piètre résultat. Le florilège des sottises publiques dans le domaine des techniques audiovisuelles vaut presque l'anthologie de celles commises en matière d'informatique.

(En 2000, la carte Sesam Vitale n'est toujours utilisable qu'auprès de peu des médecins libéraux.)

(Le rapport traite ensuite des techniques audiovisuelles sous les aspects du plan câble, des satellites TDF1 et TDF2, des normes MAC et D2MAC, des télécommunications pour le rattrapage du retard en matière de téléphone, le minitel, l'accord entre France Télécom et Microsoft et la téléphonie mobile.

D'autres errements auraient pu être cités dans le rapport, notamment l'arrêt après réussite d'une préfiguration d'internet conçue par Louis Pouzin sous le nom de [Cyclades](#), ou encore le lancement du projet Pascal-Sol par l'Inria. C'était la réécriture d'Unix en Pascal, car les universitaires et chercheurs de l'INRIA ne considéraient pas le langage C pour son utilité et sa diffusion mais pour le fait qu'on le jugeait mal construit. Pour la petite histoire, on rappellera les dépenses somptuaires de l'agence pour le développement de l'informatique créée en 1982 et dissoute pour ce motif en 1986).

ADDENDUM personnel du 26 juillet 2000


SUR INTERNET

Les aides de l'état frappent toujours deux fois. Une première fois par les impôts rendus nécessaires et une deuxième fois en empêchant par les réglementations ou en dissuadant par le confort apparent qu'elles donnent, les prétendus bénéficiaires d'adopter les bonnes solutions.

Le premier ministre a annoncé que le contribuable va devoir consacrer 4 milliards à Internet. Les mesures envisagées comprennent la fourniture d'ordinateurs aux écoles (bis repetita...). Il est prévu en plus une «grande école d'Internet». Tout cela pour stimuler le développement d'internet. Si le besoin d'une école d'internet se faisait sentir, elle aurait probablement été créée sous forme d'une entreprise. Quoi qu'il en soit, les causes du développement relativement faible

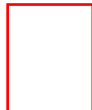
d'internet ne paraissent pas avoir été analysées. Il faudrait déjà regarder le tarif de monopole des communications locales par ce que l'euphémisme à la mode nomme «l'opérateur historique».

Dans le même temps ou presque, les ministres des finances de ce que l'on nomme le G8, réunis à Tokyo et les chefs d'états du même G8 se préparent à prendre en main internet au nom de «la justice sociale», de «l'égalité des droits» et de «la protection des plus faibles». Il s'agirait de combler le fossé numérique qui sépare ceux qui ont accès à la toile et ceux qui ne l'ont pas, en d'autres termes s'appropriier le contrôle de structures qui ne leur doivent rien.



Conservatoire National des Arts et Métiers
Architectures des systèmes informatiques
ANNEXE 10
Les choix publics en matière de techniques nouvelles
Année 2000-2001

Conservatoire national des arts et métiers
Architectures des systèmes informatiques
HORS CHAPITRES
Conclusion, première partie
Année 2002-2003



On passe la moitié de sa vie à retenir sans comprendre et l'autre moitié à comprendre sans retenir.

Antoine Rivarol

La conclusion est organisée en trois parties :

- dans la première, fichier [asi9994](#), on présente les conclusions relatives aux éléments principaux ou axes des enseignements dont les principes sont exposés dans l'annexe asi0002.
- dans la seconde on présente succinctement le statut des machines comme objets techniques fichier [asi9995](#), et on essaie de placer les machines informatiques par rapport aux précédentes, fichier [asi9996](#).
- dans la troisième, on brosse un tableau des tendances actuelles et des futurs possibles, respectivement pour le court terme, fichier [asi9997](#), le moyen terme, fichier [asi9998](#) et le long terme, fichier [asi9999](#).

Sur le contenu des notes de cours relatif aux architectures, bien que :

La bêtise consiste à vouloir conclure. Gustave Flaubert.

(mais faut-il croire ce que dit Monsieur Flaubert ?)

C 1. SUR LES IDÉES DIRECTRICES OU AXES DE PRÉSENTATION

Nous apportons en premier lieu trois jeux de conclusions relatifs aux trois axes énoncés dans l'annexe asi0002.

- Pour le premier axe,

[les considérations architecturales relèvent de la logistique et sont tournées vers sa meilleure efficacité,](#)

autrement dit : **«Il n'y a pas de démarche déductive en matière d'architectures de systèmes informatiques.»**

notre conclusion est que :

Il n'y a pas de méthode générale de conception des architectures de systèmes informatiques.

Il avait été dit dans le chapitre 2 que, sauf erreur dans les prémisses de l'analyse, il ne peut pas y avoir de théorie architecturale déduite des principes fondamentaux. Cette thèse n'a pas été prise en défaut.

Remarquons toutefois qu'il existe des moyens pour aboutir à des architectures spécialisées adaptées au traitement d'une classe de problèmes. La spécification de la classe fournit alors le critère exogène indispensable qui a deux fonctions :

- fondamentalement, réduire la généralité ou l'universalité du schéma général;
- pratiquement, mener à la mise en œuvre de moyens techniques divers.

Ce critère peut mener à définir un jeu d'instructions, c'est-à-dire une machine virtuelle. Il peut aussi mener à une architecture s'il est par lui même une structure de données, par exemple une machine SIMD.

- Pour le deuxième axe,

les méthodes et les moyens architecturaux de la logistique ne sont pas hétéroclites; ils obéissent à des règles communes,

autrement dit : **«Il y a une unité de principes et de techniques opératoires en matière de stockages et de transmissions, séparément et ensemble.»**

notre conclusion est en trois parties :

Première partie : le maquis des techniques architecturales tel qu'il est habituellement présenté est plus ou mieux ordonné qu'on peut le croire. Certes, les techniques employées ne peuvent pas être déduites d'une théorie, elles viennent donc d'ailleurs, mais elles sont très rarement de pures inventions. Le plus souvent elles sont des transferts de techniques d'autres secteurs d'activité très différents comme le livre (pagination pour un livre broché ou relié, segmentation pour un livre en rouleau), la poste pour le modèle d'interconnexion des systèmes ouverts, etc. Ces origines ont été présentées autant que possible dans le cours oral.

Nous avons considéré deux types d'activité dans le fonctionnement d'un système informatique :

- les transformations, calcul inclus;
- la logistique, entreposage ou stockage et transmission.

Nous avons introduit, dans le cours oral et dans ces notes, la logistique comme fonction principale, à côté de la transformation et du calcul qui est la fonction essentielle.

Nous en avons fait le moyen ou fil conducteur pour ordonner les techniques. Les méthodes de

gestion du stockage sur des supports différents et les méthodes des transmissions ont des caractères d'homogénéité et de cohérence.

L'unité méthodologique des techniques de stockage sur des supports différents comme la mémoire électronique et les supports magnétiques pourra être considérée après réflexion comme normale, eu égard à l'unité de la fonction et à la similitude des objectifs. On a constaté que cette unité existe aussi, pour des raisons moins immédiates, entre le stockage des données et la gestion des transmissions. Nous considérons cet apport comme le principal gain en compréhension qui justifie la logistique comme composante explicative des ordinateurs.

Pour ce qui est des transformations qui incluent le décodage des instructions et leur exécution, elles constituent l'autre domaine que la logistique ne prend pas en compte. En l'état de nos réflexions, la gestion des événements est une intrusion des phénomènes extérieurs ou intérieurs, via le temps pris comme instant. On peut la rattacher aux transformations ou encore en faire une autre catégorie. Quoi qu'il en soit, le temps comme intervalle, durée ou encore instant ne fait pas partie des matières symboliques mathématiques ou informatiques.

Deuxième partie : Les outils et les méthodes utilisables aux fins de conception, tant descriptifs que normatifs et sur les règles d'action **sont tous pauvres**, aussi bien la structuration en couches que la dualité entre temps et espace. Pour ce qui est des normes, elles régissent chacune une très petite partie du domaine en cause et sont des prescriptions destinées à dire la bonne pratique. Les essais de présentation quantitative ne guident pas mieux le concepteur.

La situation devient différente dès lors que l'on introduit un ou des critères exogènes pour traiter des problèmes particuliers, autrement dit pour concevoir des machines aptes à traiter **UN PROBLÈME** ou **UNE CLASSE** de problèmes. On entre dans le domaine de la conception systématique des objets informatiques qui fait l'objet d'autres enseignements. En termes architecturaux l'aboutissement est la machine-langage, la machine-flux de données ou encore la machine-structure de données. La pertinence de l'architecture obtenue est limitée aux seuls critères de définition. L'application d'une telle méthode produit alors une seule architecture informatique propre à traiter le problème pour la machine-langage virtuelle, alors qu'il en existe plusieurs pour les machines flux et structures de données.

À l'opposé, les outils de simulation sont puissants et de grande qualité, mais on ne simule que ce que l'on a déjà imaginé. Une simulation n'imagine pas une architecture, elle vérifie une capacité à fonctionner et éventuellement évalue des performances, comme une machine analogique simulait une réalité déjà appréhendée. Les outils de simulation sont utilisables pour vérifier la pertinence d'un choix.

Troisième partie : Sur les aspects temporels des deux composantes de la logistique.

- Le transport dans l'espace est une opération qui se déroule dans l'espace à un instant ou dans un intervalle de temps précis. Cet instant ou l'un des instants de l'intervalle porte ce qu'on nomme une synchronisation.
- Le transport dans le temps est une opération qui se déroule dans le temps et donc en un lieu précis. Elle n'apporte pas de synchronisation mais ce que l'on pourrait nommer syntoponisation. Il existe dans les bases de données relationnelles sous la forme de l'enregistrement unique.

Les multiprocesseurs à mémoire commune ont donc besoin d'outils de synchronisation ajoutés alors que, par leur définition, les multiprocesseurs à messages les contiennent.

Réciproquement, les multiprocesseurs à message ont besoin d'outils de maintien de la cohérence des données entre des lieux multiples alors que les multiprocesseurs à mémoire strictement commune les contiennent.

Les multiprocesseurs mixtes, soit parce qu'ils combinent les deux modes, soit parce qu'ils ne relèvent pas d'un mode pur sont dans des situations intermédiaires.

- Pour le troisième axe,

la tendance profonde en matière architecturale est de réaliser la plus grande simultanéité des traitements par un parallélisme à tous les niveaux possibles,

autrement dit : «Les concepteurs introduisent la simultanéité par le parallélisme architectural à tous les niveaux envisageables.»

notre conclusion est que :

La mise en parallèle des composants architecturaux est une tendance lourde qui se poursuivra.

La mise en parallèle des bits paraît achevée, car il ne semble pas utile d'envisager des mots de données de plus de 64 bits, sauf le cas de registres multimots qui relèvent du parallélisme entre données.

La mise en parallèle des outils de décodage et d'exécution des instructions par pipeline est irréversible. Le nombre d'étages sera limité par les conséquences des mauvaises prédictions de branchement. La mise en parallèle des instructions (VLIW) n'est pas assurée du succès. Le choix entre le superscalaire et le VLIW n'est pas fait. Peut-être y aura-t-il plus convergence qu'affrontement.

La mise en parallèle des données (MMX chez Intel, 3Dnow! chez AMD) qui est une version réduite de l'architecture SIMD sera guidée par la nature des données manipulées.

La mise en parallèle des processeurs, sous la forme de quelques processeurs fonctionnant avec une mémoire commune paraît acquise (architecture symétrique). Quatre processeurs sur une même carte est banal, quatre processeurs sur la même puce deviendra possible. On peut aussi attendre une spécialisation des processeurs sur une même puce, par exemple l'un général, l'autre spécialisé dans la gestion des appels au système, comme l'Univac 1108 avait un processeur affecté aux entrées et sorties. Les systèmes massivement parallèles paraissent encore et peut-être pour longtemps des vues de l'esprit pour une utilisation générale. Toutefois, des progrès dans les systèmes d'interconnexion pourraient rendre viables dix à quelques centaines de processeurs.

C 2. QUELQUES AUTRES CONCLUSIONS

C 2.1 Sur les deux machines physique et symbolique.

Cette distinction apporte des vues différentes malgré l'imprécision de la notion de machine

symbolique qui possède plusieurs niveaux d'abstraction.

C 2.2 Sur la conservation des idées.

Un fait a été observé à plusieurs reprises : un type de dispositif est inventé, son application est faite ou ajournée ou abandonnée selon l'évolution des techniques du moment.

Nous l'avons vu pour la microprogrammation, c'est aussi le cas des [circuits asynchrones](#). Ce pourra être le cas des jeux d'instructions qui sont devenus complexes, revenus à la simplicité il y a quelques années et pourront resurgir encore plus complexes dans l'avenir.

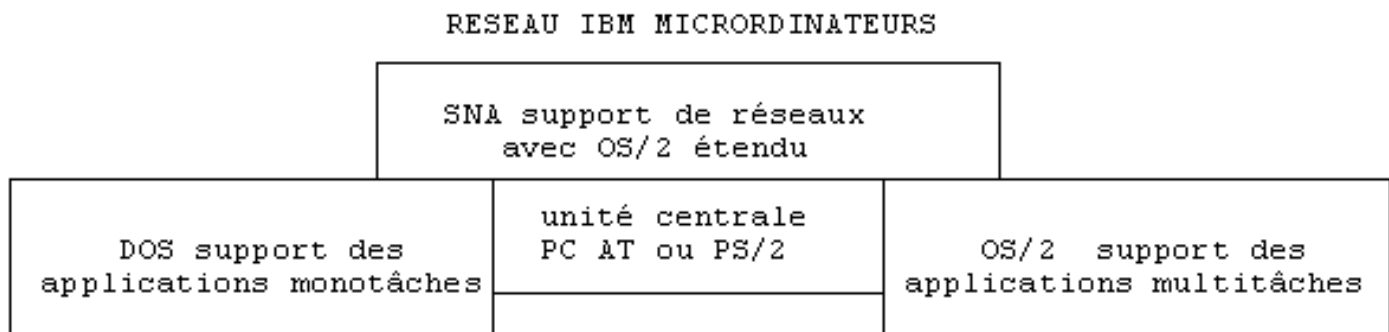
Le lecteur pressé peut penser qu'il n'y a pas lieu de modifier ses habitudes et que si une technique revient, il suffira de l'apprendre au bon moment. Nous pensons au contraire que la mémoire du passé est importante et que la conservation des procédés mis en attente doit être faite par chacun. **«Qui ne connaît pas le passé est condamné à le revivre.»**

C 2.3 Sur d'autres acceptations du mot architecture

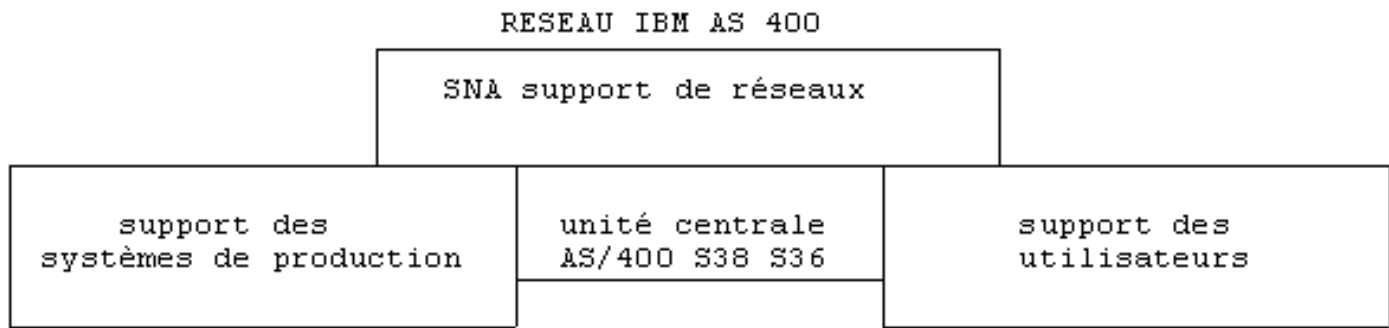
On nomme aussi architecture l'ensemble des règles qui précisent le fonctionnement d'un ou plusieurs matériels, incluant les systèmes d'exploitation, ainsi que leurs moyens de communication externe, sans plus se soucier de leur constitution interne. Cette acception est utilisée par IBM, dans des documents qui relèvent aussi bien du technique que de la mercatique. On donne ci-après trois présentations de ce qu'elle recouvre.

On lit de la même façon depuis deux ans l'expression «Architecture client-serveur». Il s'agit ici aussi d'un mode d'utilisation de plusieurs machines reliées entre elles.

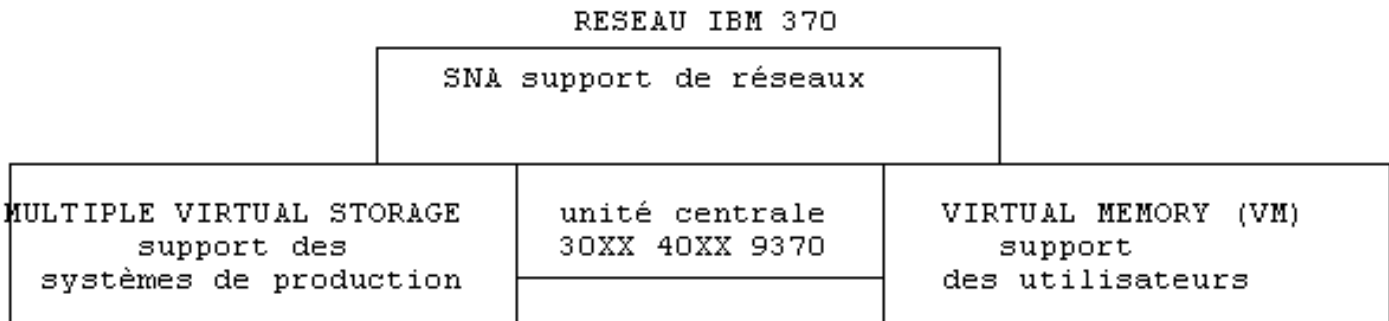
C 2.3.1 Les micro ordinateurs et leur réseau, IBM



C 2.3.2 Les systèmes petits et moyens, architecture AS 400, IBM



C 2.3.3 Les grands systèmes, architecture 370, IBM



Le fédérateur de ces trois architectures est manifestement le support de réseau SNA installé sur les trois, comme il l'est sur trois produits complémentaires : S/88 à tolérance de pannes, S/1 dit «à tout faire» spécialisé dans la communication, et 6150 sous AIX (variante d'unix) poste de travail pour la CAO et la DAO.

C 2.3.4 Sur l'architecture client-serveur

Cette expression est tout autant le reflet d'un changement de mode de fonctionnement de la micro informatique qu'un terme d'appel publicitaire. Tous les systèmes commerciaux deviennent «client-serveur» comme ils sont déjà «puissant», «souple», «convivial», etc. Ces mots sont à la mode.

Plus techniquement l'architecture client-serveur consiste à réaliser au moyen de postes en réseau ce qui existe depuis l'arrivée de la multiprogrammation.

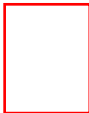
Par la multiprogrammation chaque poste connecté dispose de ses programmes et ceux-ci ont accès aux ressources communes, une base de données et son moteur par exemple.

En micro informatique isolée, chaque utilisateur dispose pour lui-même de la totalité des ressources, bases de données incluses logées dans sa machine.

En micro informatique en réseau, les accès aux bases de données par chaque poste nécessitent la lecture complète de la base ou d'un extrait cohérent, via le réseau. Les opérations sont ensuite faites localement jusqu'à l'émission de l'ensemble s'il y a eu mise à jour. Bien entendu, en cas de mise à jour, la totalité de la base est fermée en mise à jour à tous les autres postes. Les quantités d'information qui transitent sur le réseau sont considérables.

Dans le mode client-serveur, le réseau existe, il portera la requête. Celle-ci sera

interprétée dans le serveur qui retournera très exactement l'information demandée. La voie est alors ouverte à l'installation d'une machine puissante jouant le rôle de serveur, capable d'interpréter les requêtes et faire les extractions et mises à jour rapidement. Elle est également ouverte à l'existence de plusieurs serveurs contenant des morceaux d'une même base ou des bases différentes. Chacun des serveurs est sollicité en tant que tel en fonction de la demande du client.



Conservatoire national des arts et métiers
Architectures des systèmes informatiques

HORS CHAPITRES

Conclusion

Année 2002-2003



Conservatoire national des arts et métiers
Architectures des systèmes informatiques
HORS CHAPITRES
Conclusion, deuxième partie
Année 2002-2003



Deux populations d'experts [...] se tournent le dos, les instruits incultes et les cultivés ignorants.

Michel Serres et al. [[annexe asi0019](#)]

SUR LES MACHINES

C 3. SUR LES MACHINES EN GÉNÉRAL ET LEURS RELATIONS AVEC LA NATURE

Ce paragraphe relativement long prend la question des machines informatiques très en amont, à partir de ce que l'on nomme la nature. On examinera succinctement ce qui constitue la nature, objets de la physique et êtres vivants. On s'intéressera ensuite aux machines par le biais de ce que disent certains historiens des sciences. Pour cela on présentera plus particulièrement la notion de paradigme de Thomas Kuhn.

C 3.1. La nature

La nature, mot souvent écrit sans bonne raison «Nature», contient la *matière* ou *énergie*, les *êtres* et, spécialement pour notre propos, les *machines* que l'homme fabrique. Nous écrivons quelques mots seulement de l'énergie-matière et des êtres, objets de toutes les attentions des physiciens, des biologistes et des philosophes.

C 3.1.1 Les objets d'étude de la physique

Après Descartes ils ont été la matière et le mouvement et plus récemment l'énergie.

Bernard d'Espagnat lors d'un colloque à l'académie des sciences morales et politiques a dit sur ce sujet :

«La liste qu'il dresse [Descartes dans les Principes de la philosophie] des "notions claires et distinctes qui peuvent être en notre entendement touchant les choses matérielles" ne comporte que celles "des figures, des grandeurs et des mouvements, et des règles suivant lesquelles ces trois choses peuvent être diversifiées l'une par l'autre", et il nous dit expressément qu'il faut "que toute la connaissance que les hommes peuvent avoir de la nature [soit] tirée de cela seul". Cette conception, qu'on appelle le mécanicisme, revient, on le voit, à considérer que, du moins dans le domaine de "l'étendue", autrement dit de la matière, tout, en fin de compte, est descriptible au moyen des concepts de la liste en question, dont on remarque tout de suite que tous ses éléments sont des concepts familiers, des concepts dont, depuis la nuit des temps, l'homme a toujours disposé.»

On appréhende la matière, l'énergie et leurs mouvements par l'expérience sensible parfois trompeuse. On construit des lois pour calculer leurs relations et leurs évolutions, que l'on pense connaître de mieux en mieux. On s'interroge indéfiniment sur leurs attributs.

accordent à la matière.

L'étude de l'énergie en tant que telle (thermodynamique) a commencé au XIX^e siècle. Mayer, Colding, Joule et Helmholtz, chacun séparément, avaient établi l'équivalence entre la chaleur et le travail (première loi) entre 1842 et 1847 [KUH62]. Rodolphe Clausius a en mains la mise en forme analytique par Clapeyron du mémoire de Carnot «*Réflexions sur la puissance motrice du feu et les machines propres à développer cette puissance*», publié à compte d'auteur en 1824 en 200 exemplaires. Il écrit en 1850 un premier texte sous le nom de «*théorie mécanique de la chaleur*», puis sous une forme définitive en 1865, la deuxième loi de la thermodynamique. Sadi Carnot avait utilisé la théorie du calorique dans son mémoire, ce fut une cause d'oubli.

L'équivalence entre matière et énergie a été établie par la théorie de la relativité générale dans les années 1910. À partir de là, le mouvement, comme l'espace, ne sont plus absolus mais relatifs. La simultanéité n'est plus une notion évidente. Les synthèses a priori de Kant s'évanouissent.

Dans le même temps, la connaissance que nous avons de la matière est bouleversée par la mécanique quantique, la notion de localité n'est plus évidente. Les philosophies ou doctrines matérialistes fondées sur la permanence, l'éternité de la matière sont privées de leurs socles d'évidences.

Sven Ortoli et Jean-Pierre Pharabod, [ORT84] auteurs du «*Cantique des Quantiques*» écrivent : «*La science des XVIII^e et XIX^e siècles avait abouti au triomphe du matérialisme mécaniste, qui expliquait tout par l'agencement de morceaux de matière minuscules et indivisibles – agencement réglé par diverses forces d'interaction qu'ils exerçaient entre eux. Cette vision assez primitive, à laquelle se tiennent encore la plupart des biologistes, avait pour conséquence l'inutilité des religions et des philosophies, qui font appel à l'existence d'entités non matérielles. Le fait que ces morceaux de matière se soient révélés n'être en réalité que des abstractions mathématiques non locales – c'est-à-dire pouvant s'étendre sur tout l'espace et de plus n'obéissant pas au déterminisme – a porté un coup fatal à ce matérialisme classique. Certes, le matérialisme est encore possible mais c'est un matérialisme quantique, qu'il faudrait appeler "matérialisme fantastique" ou "matérialisme de science-fiction"».*

C 3.1.2 Les êtres.

Le statut des êtres a été lui aussi considéré très diversement. Au XVIII^e siècle la [conception mécaniste](#) que nous avons déjà mentionnée pour la physique a été étendue aux êtres par Descartes, Malebranche, puis La Mettrie, Diderot et Lavoisier entre autres. Elle a été développée par les scientifiques et les matérialistes du XIX^e siècle. Dans cette conception, rien ne distingue un être vivant d'une machine très complexe, Descartes a parlé le premier d'animaux-machines. Pour un mécaniciste, le fonctionnement des êtres relève de la seule physique de la matière. Ils ont été nommés positivistes au XIX^e siècle, matérialistes, et parfois [réductionnistes](#) (dans le sens d'une attitude philosophique générale de réductionnisme ontologique). Cette conception imprègne encore fortement notre conception du monde. Claude Tresmontant [TRE63] écrit en 1953 : «*Le mécanisme de type cartésien a eu des conséquences incalculables et désastreuses aussi bien en biologie qu'en médecine et en psychologie. Considérer l'organisme humain comme une machine dont on pourrait traiter les pièces séparément, et, corrélativement, dissocier le psychique du somatique, telles sont les deux erreurs mortelles dont les conséquences se font encore sentir aujourd'hui.*» Comme le dit le neuropsychiatre Boris Cyrulnik : «*C'est Descartes qui nous a fait ce coup ! [...] L'homme que vous avez devant vous est entier, c'est une totalité*» (émission sur Europe1 le 30 janvier 2003 à 9h30).

Le mécanicisme a pratiquement toujours été déterministe, il portait l'espoir de tout savoir et tout prédire. Laplace a écrit dans son «*Essai philosophique sur les probabilités*» de 1814 [LAP14]: «*Une intelligence qui, pour un instant donné, connaîtrait toutes les forces dont la nature est animée et la situation respectives des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre toutes ces données à l'analyse, embrasserait dans la même formule les mouvements des plus grands corps de l'univers et ceux du plus léger atome. Rien ne serait incertain pour elle et l'avenir, comme le passé serait présent à ses yeux. L'esprit humain offre, dans la perfection qu'il a su donner à l'Astronomie, une faible esquisse de cette intelligence*»

Ces trois phrases ont été abondamment citées car elles contiennent l'essentiel de cette vision du monde. «*Cette idée a exercé une fascination*

profonde sur des générations de personnes à l'esprit scientifique...» Friedrich August Hayek dans **The counter-revolution of science - Studies of the abuse of reason**, *Economica* (dans les années 40).

Le lecteur peut déduire au moins deux conséquences très immédiates du mécanicisme déterministe :

- la capacité de choix et la liberté de l'individu n'existent pas puisque tout l'avenir est connaissable, le déterminisme est absolu;
- le futur est entièrement conditionné par le passé, il ne peut donc pas contenir de nouveauté, le monde ne peut être qu'éternel.

De plus le temps est alors un résidu puisque les lois de la physique ne peuvent être que réversibles, ce qui est vrai dans la physique newtonienne et dans la physique quantique.

À l'opposé, on a appris beaucoup plus récemment :

- que le monde a commencé (le big bang);
- qu'il est en évolution (cosmologie, géophysique);
- que l'aléatoire et l'incertitude existent en physique et dans les modèles mathématiques (les chaos, déterministes ou non, des phénomènes macroscopiques);
- l'incertitude limite fondamentalement la connaissance que l'on peut avoir des phénomènes quantiques;
- que les propositions indécidables existent en mathématiques.

Le futur n'est pas entièrement conditionné par le passé.
La conviction de Laplace n'est plus de mise.

Cette conviction provient presque directement des conceptions des grecs et d'autres anciens : le monde est fixe, les modifications que l'on y voit ne sont que temporaires, tout redevient un jour identique au passé. Cette conception conduisait soit au temps absolu et à l'espace absolu de Newton, soit si l'on rejetait ces absolus, aux doctrines de l'éternel retour du monde cyclique (métaphysique brahmanique et éternel retour nietschéen). Elle était à l'opposé des deux métaphysiques juive et chrétienne.

Le mécanicisme fleurit aussi en informatique spécialement dans ce que l'on nomme parfois «l'intelligence artificielle dure». La recherche de Turing sur *«Les machines peuvent-elles penser ?»* en est un exemple. Pour éclairer ce point en montrant comment on le fonde, nous reprenons et commentons entre [...] une partie d'un excellent texte d'Andrei Popescu-Belis,

<http://www.polytechnique.fr/eleves/binets/xpassion/articles/xpl2turi.html> à propos de la thèse de Church-Turing :

«Un premier pas que l'on peut franchir est celui des "processus logiques" chez les humains. En effet, la plupart [dernière concession de l'auteur] des assertions mathématiques sont formalisables en termes d'algorithmes, voire de théorèmes sur les entiers (Gödel) - donc finalement se ramènent à la question de savoir si un nombre possède ou non telle propriété, et à un algorithme pour y répondre. Ceci n'est en fait qu'un raccourci de langage pour rendre plus concrète la notion de raisonnement mathématique, mais l'essentiel est que celui-ci est formalisable au moyen d'algorithmes du langage "standard". [Le fonctionnement varié du cerveau humain est ici discrètement, sans le mentionner aucunement, réduit à des processus logiques et ensuite à des processus algorithmiques. La restriction "la plupart" du début de l'alinéa a disparu.]

Or, les humains effectuent couramment de tels "raisonnements". [Voilà que la pente est glissante.] *La thèse de Church-Turing peut alors être formulée comme suit: "Tout processus mental qui trie les nombres en deux classes peut être simulé par un programme". Pour éliminer des "méthodes" comme l'intuition géniale, et fiable, on peut restreindre la thèse aux processus communicables à un autre individu, par exemple. Cette thèse aussi semble vraie jusqu'à présent: la plupart des mathématiciens s'accordent à dire que leurs aptitudes ne dérivent pas d'une intuition transcendante des mathématiques mais de l'aptitude à enchaîner les raisonnements symboliques.* [Les humains sont-ils limités à faire des tris ?]

Il n'y a pas de raison de séparer les processus mentaux "mathématiques" des autres formes de raisonnement- on voit mal pourquoi il y aurait une frontière nette [ce n'est pas parce qu'on voit mal quelque chose qu'elle n'existe pas !]; ceci nous conduit à la formulation réductionniste de la thèse de Church-Turing: "Tous les processus mentaux peuvent se ramener à des programmes". [Et hop, le pas est franchi sans démonstration ni argument] Personne ne prétend avoir démontré cette thèse, qui semble d'ailleurs très problématique. Nous avons montré le cheminement qui y conduit, et qui suggère que cette hypothèse n'est pas entièrement sans fondement. Elle fournit aujourd'hui le but plus ou moins avoué d'une nouvelle discipline: l'intelligence artificielle. [Cette restriction paraît n'être que de précaution, elle arrive bien tard, après les efforts faits pour convaincre le lecteur du contraire. Mise en tête du §, elle aurait eu plus de force. Il s'agit là non pas d'une démonstration scolastique mais d'une tentative faite pour entraîner l'adhésion à une thèse, le moyen-âge où l'on s'efforçait de démontrer a cédé la place au discours persuasif de la [Renaissance](#). La conclusion est redoutable : une discipline, c'est-à-dire une part de la science, a une hypothèse pour but !]

Note sur le réductionnisme

Très brièvement le réductionnisme consiste à penser ou croire qu'il faut réduire les aspects divers et multiples de la réalité au plus petit nombre d'éléments suffisant pour faire apparaître la connaissance. Le réductionnisme a plusieurs aspects dont les deux suivants.

L'aspect systémique quand on affirme que la connaissance des phénomènes ne peut être atteinte qu'en réduisant leurs descriptions multiples à un nombre toujours plus restreint de principes, de concepts, de théories et de lois. Les grandes théories et les grandes lois sont de ce type : gravitation de Newton puis d'Einstein, unification de l'électricité et du magnétisme par Maxwell. Cet aspect est une attitude normale en matière de sciences.

L'aspect ontologique ou philosophique quand, voulant définir la réalité en tant qu'être, on la considère comme une hiérarchie de niveaux, chacun possédant ses constituants propres jusqu'à un niveau fondamental si l'on y arrive. Les éléments d'un niveau sont alors explicables par les caractères du niveau inférieur. Cet aspect est largement controversé dans la mesure où l'on en vient, par exemple, à devoir expliquer le vivant par des propriétés plus ou moins cachées de la matière. La mention que [nous avons faite précédemment](#) portait sur le réductionnisme ontologique et non comme pratique scientifique.

La partie relative à la logistique des présentes notes participe du réductionnisme systémique au contraire de la proposition relative à la machine double.

Fin de la note.

Les connaissances actuelles en matière de biologie ne sont pas interprétées ni même reconnues de façon unanime. Certains biologistes tiennent à la conception mécaniste. La croyance en l'univers éternel ayant disparu, certains expliquent la nouveauté qui apparaît dans le monde par des hasards successifs (Jacques Monod dans «Le hasard et la nécessité»). Ces hasards remplacent le Dieu arbitraire de Newton et le Dieu initial de Descartes. D'autres présentent des objections fortes à cela.

- La première est l'immense difficulté à donner une explication mécanique de l'apparition de la vie, même dans ses manifestations les plus simples. La vie met en œuvre des molécules d'une telle complexité que leur assemblage par le hasard n'est ni prouvable ni réfutable quelle qu'ait pu être la situation physico-chimique de la terre il y a trois à quatre milliards d'années. On n'a pas pu la recréer en laboratoire. Une autre façon de rechercher ou d'expliquer la naissance de la vie consiste à chercher l'origine de l'information correspondante. Ce n'est ni plus ni moins simple.
- La seconde est la distinction bien établie mais qui est mal ou pas comprise entre la matière et le vivant. C'est une question centrale des sciences. La matière (matière et énergie) se comporte en accord avec [la seconde loi de la thermodynamique](#) établie au XIX^e siècle, loi de Carnot pour les français, de Clausius pour les anglo-saxons, dite encore loi de l'entropie croissante. À l'opposé, on constate que si chaque être vivant va vers la dégradation et la mort, collectivement les êtres vivants tendent vers une complexité plus grande. L'information qu'ils contiennent et l'information nécessaire à leur édification contenue dans le génome (chaîne d'ADN) est de plus en plus grande de l'amibe aux animaux supérieurs. Les hasards peuvent-ils l'expliquer de façon convaincante ? On trouvera plus loin l'incidence de cette loi en informatique.

Enfin et pour ce qui nous concerne,

Une machine	Un être vivant
a des constituants fixes;	est constitué de cellules qui se reproduisent;
reçoit et consomme de l'énergie pour produire ce pour quoi elle a été construite par l'homme;	reçoit et consomme de l'énergie pour fonctionner selon son plan propre;
ses composants sont figés; ils s'usent et sont remplacés.	les cellules se renouvellent constamment par <i>assimilation</i> de molécules nouvelles.

Ceci peut suffire, au moins à première vue, à faire un sort aux abus de langage tels que cerveau électronique, virus, machines pensantes, etc.

C 3.1.3 Note intermédiaire sur les lois de la thermodynamique.

La thermodynamique a deux lois fondamentales établies pour un système dit isolé :

- une loi de **conservation** ou 1ère loi, (Mayer, Colding et Joule)
 - *l'énergie interne d'un système isolé est invariante;*
- une loi de **transformation** ou 2ème loi, (Carnot et Clausius)
 - *les changements spontanés d'un système isolé font croître une grandeur nommée entropie.*

La deuxième loi a d'autres formulations comme «*l'énergie tend spontanément à s'écouler à partir de sa forme concentrée en un lieu*», ou encore «*les systèmes physiques tendent par leur fonctionnement vers l'uniformisation*». Elle a un statut particulier en ce qu'elle est une inégalité et non une égalité comme l'est la première et la plupart des lois de conservation de la physique. Elle dit que dans un système isolé, l'énergie calorifique ne peut pas être transformée totalement en énergie mécanique. La partie non transformée est un résidu d'énergie calorifique à température plus faible. La quantité totale d'énergie est conservée mais la température du résidu a été dégradée. Cette transformation est irréversible. La dégradation est mesurée par la grandeur nommée entropie.

La difficulté intrinsèque à cette loi tient à l'expression : système isolé. Les systèmes isolés par eux mêmes ou comme sous ensembles de systèmes isolés sont difficiles à cerner. C'est le cas pour les êtres vivants et pour l'univers pris dans sa totalité. La thermodynamique distingue :

- les systèmes proprement isolés, il n'y a pas de transfert de matière ni d'énergie à travers leurs frontières;
- les systèmes adiabatiques, il n'y a pas de transfert de chaleur ni de matière à travers leurs frontières, des transferts d'autres énergies sont acceptés;
- les systèmes fermés ou clos, seul le transfert de matière est interdit;
- les systèmes ouverts, tous les transferts d'énergie et de matière sont possibles.

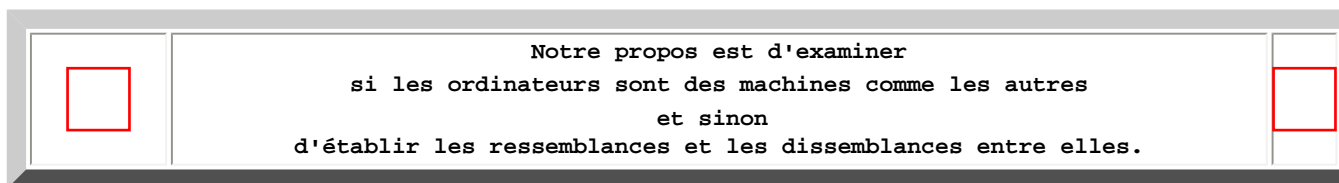
Une grande question pendante est : est-ce que l'univers est un système isolé ? Si oui il se dégrade continûment, les étoiles s'éteindront et l'univers deviendra uniformément froid, et s'il n'est pas isolé ? eh bien, on ne sait pas.

C 3.2 Les machines.

Les machines sont évidemment de nature mécanique ne serait-ce que parce que nous ne fabriquons pas du vivant même si nous le transformons. En font partie les mécanismes au sens usuel, les machines thermiques, électriques, chimiques et autres et plus généralement les fabrications humaines.

Les machines sont définies et construites par l'homme selon des plans et pour des finalités précises, EN RESPECTANT à tout le moins les **lois de la physique** relatives aux phénomènes utilisés dans leur domaine de validité et en APPLIQUANT ces lois quand on les connaît. On ne peut pas construire une machine en violation d'une loi valide car si on le pouvait, la loi ne serait alors plus valide. L'exemple le plus manifeste de machine impossible est le moteur fournissant de l'énergie sans dépense en opposition aux lois de la thermodynamique.

La **validité des lois** repose, au moins en partie, sur leur caractère prédictif à partir duquel on peut concevoir ces machines utilisables. Ceci **n'implique pas** que la loi exprime le **pourquoi** du phénomène ou de la transformation.



Commençons par quelques considérations utiles sur la physique, ses théories et ses lois. Nous n'entrons pas dans la dispute philosophique entre idéalistes et réalistes sur le degré de réalité de la physique et de ses modèles et sur le statut des théories et de leurs lois. Suivant en cela nombre de physiciens, nous nous limitons à considérer les théories et les lois que l'on en déduit pour leur utilité à prévoir des résultats et pour la précision de leurs prédictions. Pour cela, il est utile d'examiner l'évolution des lois elles-mêmes.

La physique est en progrès, souvent cumulatifs, plus rarement destructifs. Pour l'historien des sciences Thomas Kuhn [\[KUH83\]](#), l'apparition d'une théorie nouvelle constitue le changement de ce qu'il nomme **paradigme**. Une science nouvelle s'établit fermement quand on a construit son paradigme. Très brièvement le paradigme est la façon commune aux hommes d'une même science de considérer et d'interpréter la partie du monde dont ils s'occupent. Selon Isabelle Stengers et Judith Schlanger [\[STE88\]](#) : *«Une science, lorsqu'elle se fonde, ne se détache pas sur un fond d'erreur, elle découvre un mode de conceptualisation capable de créer l'unanimité»*. C'est une vision d'une partie du monde, accompagnée d'un jeu de règles pour appliquer cette vision. Certains philosophes nommés *«déconstructeurs»* ou «postmodernes» en ont tiré l'affirmation selon laquelle les scientifiques sont gouvernés par une psychologie de masse et non par la rationalité. Que les scientifiques passent d'un paradigme à un autre au cours de révolutions scientifiques montrerait pour eux qu'il n'existe pas de monde réel que la science doit décrire. Chaque nouveau paradigme créerait son propre monde. (extrait de *John R. Searle*, Université de Californie, Berkeley, **Rationality and Realism, What is at Stake ?** paru dans *Daedalus*, automne 1993, pp. 55-83.) Cette vision reprend au fond l'affirmation première des philosophes de l'Un, brahmanique, spinoziste ou autre.

Les [historiens des sciences](#) ne sont pas unanimes pour suivre Thomas Kuhn qui relève plutôt du groupe idéaliste (au sens technique du terme, c'est-à-dire donnant la primauté à l'idée). **MacKenzie** qui est plus empiriste, définit une culture technique comme l'ensemble des connaissances, des idées, des pratiques et des outils basés sur elle. Pour lui, il y a des évolutions plus ou moins rapides et non des changements brutaux. Une histoire des sciences qui se veut histoire des concepts pourrait aussi être une histoire linguistique des sciences qu'établirait un tenant de la philosophie analytique. On verrait alors que les révolutions scientifiques sont chacune liée à un nouveau langage, lui-même solidaire de catégories neuves (Galilée), que leur progrès dépend de l'amélioration de leur nomenclature (Lavoisier), de leur vocabulaire (Linné), qu'un changement, voire un «viol» linguistique (Riemann), peut entraîner un bouleversement conceptuel générateur d'extension théorique (géométries non euclidiennes), que leurs théories s'expriment dans la structure de leurs équations (Maxwell, Poincaré), avant même que leur axiomatisation n'achève de révéler la nature linguistique des problèmes eux-mêmes (court extrait de l'article *«Philosophie analytique»* par Francis Jacques in *Encyclopædia Universalis*).

[Bertrand Gille](#) part de ce que l'histoire de l'humanité est scandée par la succession de «systèmes techniques», combinaison de quelques technologies fondamentales. Il illustre cette idée en décrivant les systèmes techniques de la pierre taillée, de l'industrie antique qui utilisant les esclaves, de l'industrie manufacturière à base d'énergie éolienne et de moulins, puis de la machine à vapeur, de la mécanique et de la chimie qui ont dominé le XIXème siècle et la première moitié du XXème, enfin de l'électronique et de l'informatique, le système technique d'aujourd'hui.

Les systèmes techniques ne déterminent pas l'histoire, ils déterminent le possible pour la production des biens et services et donc pour la satisfaction des besoins. Ils conditionnent le possible économique et, par lui, influencent l'histoire.

Nous ne ferons que mentionner un des principaux déconstructeurs, [P. K. Feyerabend](#) (1924-1994) qui soutenait que la science est faite de nos désirs et que nous «produisons» la science que nous voulons avoir (voir aussi [H. Guénin-Paracini](#)).

Revenons néanmoins au paradigme de Kuhn. Pour éclairer cette notion prenons l'astronomie, une des sciences les plus anciennes.

Le **paradigme de Ptolémée** consistait à admettre que la terre, que l'on savait ronde depuis au moins le calcul de son rayon par Eratosthène, est le centre du monde, en accord les principes du mouvement de la physique d'Aristote.

Il a été remplacé par le **paradigme de Copernic** où le soleil était le centre du monde.

Lui même a été renouvelé par **Newton** et remplacé par **Einstein**.

En astronomie, chaque paradigme est établi sur une théorie : les mouvements circulaires de Ptolémée, les mouvements elliptiques de Képler et Newton, les mouvements selon les géodésiques de l'espace-temps d'Einstein. Ces théories fournissent les règles de calcul pour prévoir les positions des planètes et les éclipses. Les astronomes vérifient que ces prévisions sont en accord avec les expériences que l'on peut faire avec les moyens du moment. Un seul fait astronomique connu dans le XIX^e siècle n'était pas expliqué par la théorie de Newton, c'était un écart faible entre prévision et expérience relatif au périhélie de Mercure. La théorie d'Einstein l'a intégré.

Pour Kuhn, les connaissances évoluent entre la progression et la stagnation entre deux changements de paradigme. Il nomme ces temps les périodes de science normale. Il analyse comment les paradigmes sont renouvelés. On a des exemples nombreux de ces renouvellements.

À la fin du XVIII^e siècle la théorie du phlogistique, le fluide qui faisait qu'un corps est combustible, a été remplacée par la combustion à l'oxygène.

Plus près de nous, après de nombreuses vicissitudes, la théorie de l'éther a disparu quand la théorie de la relativité a été fondée au début du XX^e siècle. L'existence de l'éther était jugée indispensable pour de nombreuses raisons dont la justification des actions à distance qui ne se concevaient alors que par des contacts successifs. On lira ci-dessous les caractéristiques de l'éther comme explication de l'aspect ondulatoire de la lumière.

Extrait de l'article OPTIQUE

ENCYCLOPEDIE MODERNE, Tome 22, 1850

par M. A. DUPONCHEL, médecin en chef de l'école Polytechnique.

505

OPTIQUE

506

ment. Suivant Newton, c'est une file de molécules ayant toutes un mouvement de transport et se succédant sans interruption.

Ces deux hypothèses comptent chacune en leur faveur des autorités d'un grand poids. Cependant, de nos jours, la théorie des ondulacions est généralement adoptée, comme expliquant les faits d'une manière complète

des actions qu'il éprouve de la part des atomes pesants. En vertu de ces forces, l'éther est uniformément répandu dans tout l'espace vide de matière pondérable; sa densité est constante et son élasticité est la même dans tous les sens. Dans un espace occupé par un corps solide, liquide, ou gazeux, l'éther peut avoir une densité plus grande ou plus petite que

et sans nécessiter aucune de ces hypothèses contradictoires que la théorie de l'émission est forcée d'admettre; elle établit un lien naturel entre les phénomènes en apparence les plus dissemblables; enfin, comme preuve irrécusable de sa réalité, elle a devancé la physique expérimentale en lui indiquant plusieurs faits que celle-ci n'avait pu soupçonner, et qui depuis ont été complètement vérifiés.

Interférences.

Entre autres phénomènes qui viennent à l'appui de l'hypothèse des ondes lumineuses, nous citerons celui des *interférences*. Ce fut le docteur Young qui démontra le premier que dans certaines circonstances *la lumière ajoutée à la lumière produit de l'obscurité*, et qui fit connaître la loi de ce singulier phénomène. En voici l'énoncé : deux rayons partis d'une source commune, et se rencontrant sous des directions peu inclinées entre elles, se détruisent mutuellement, quand il y a entre les longueurs des chemins qu'ils ont parcourus un certain nombre impair de fois une certaine quantité *d* donnée par l'expérience. Quand la différence de marche, au contraire, est nulle, ou contient un certain nombre pair de fois cette même quantité, les deux rayons se renforcent mutuellement, et la lumière résultante est à son *maximum*.

dans le vide, et son élasticité suit les mêmes variations que celle des corps pondérables, c'est-à-dire qu'elle est constante dans les gaz, les liquides et les solides homogènes non cristallisés, mais qu'elle varie avec la direction dans les cristaux dont la forme primitive n'est point un polyèdre régulier.

2° Les corps lumineux vibrent comme les corps sonores, mais avec beaucoup plus de rapidité. Les vibrations de leurs particules sont communiquées à l'éther, se propagent dans ce fluide, et donnent lieu à des ondes qui produisent la sensation de la lumière. Des vibrations plus ou moins rapides occasionnent des ondes lumineuses plus ou moins larges, d'où résulte la sensation des différentes couleurs.

Transmission de la lumière.

Toute ligne partant d'un corps lumineux, et que suit la lumière en se propageant, se nomme *rayon de lumière*. Dans l'hypothèse de l'émission, ce rayon ne peut être qu'une ligne droite; et en effet cette conclusion se vérifie le plus généralement, car il arrive presque toujours qu'on ne peut voir un corps lumineux quand il existe entre ce corps et l'œil, sur la ligne droite qui les joint, un milieu opaque, ou au travers duquel la lumière ne peut se propager.

Il est cependant telle circonstance où la

On peut comparer, jusqu'à un certain point, les interférences des ondes lumineuses à ce qui se passe dans une eau tranquille quand on y jette quelque corps solide. Il n'est personne qui n'ait remarqué que quand deux groupes d'ondes se croisent sur la surface de l'eau il y a des points de rencontre où l'eau reste immobile si les deux systèmes d'ondes sont à peu près de la même force ; tandis qu'il y en a d'autres où les ondes se renflent par leur réunion.

Bien que les recherches de Young aient mis hors de doute les faits qui servent de base à la théorie des ondes lumineuses, cependant elle doit ses progrès les plus importants aux travaux plus récents de Fresnel. Les principes de cette théorie peuvent se réduire aux deux suivants :

1° Il existe dans tout l'espace, et même entre les particules des corps, un fluide éminemment élastique auquel on donne le nom d'éther. L'état statique de ce fluide dépend de la répulsion qu'il exerce sur lui-même, et

lumière semble marcher en ligne courbe quoique dans un milieu homogène. Ce fait remarquable constitue une objection d'une grande force contre la théorie de l'émission, car avec elle on ne peut ni l'expliquer ni même en concevoir la possibilité ; tandis qu'avec la théorie des ondes on s'en rend parfaitement compte. Cependant, les cas où la lumière suit une ligne courbe étant tout à fait exceptionnels, on admet généralement que dans le même milieu la lumière se propage en ligne droite, c'est-à-dire que les rayons lumineux sont des lignes droites. Cet axiome est du reste des plus faciles à démontrer expérimentalement : si l'on pratique une petite ouverture au volet d'une chambre obscure, la lumière, en pénétrant dans la chambre, éclaire tous les corps qui sont sur son passage, et donne naissance à une ligne brillante rectiligne ; mais si sur la ligne droite qui joint le point lumineux à l'œil il se rencontre un corps opaque, ce point cesse d'être éclairé, et par conséquent d'être aperçu.

Aujourd'hui encore il subsiste quelque incertitude sur le sujet, soutenue notamment par Maurice Allais.

Sur tout cela voyons l'opinion d'un très bon connaisseur, J. de Neumann [\[NEU63\]](#) :

«... there have been within the experience of people now living at least three serious crises... There have been two such crises in physics--- namely, the conceptual soul-searching connected with the discovery of relativity and the conceptual difficulties connected with discoveries in quantum theory... The third crisis was in mathematics. It was a very serious conceptual crisis, dealing with rigor and the proper way to carry out a correct mathematical proof. In view of the earlier notions of the absolute rigor of mathematics, it is surprising that such a thing could have happened, and even more surprising that it could have happened in these latter days when miracles are not supposed to take place. Yet it did happen.»

... il y a eu dans l'expérience des hommes qui vivent aujourd'hui au moins trois crises sérieuses... Il y en a eu deux en physique--- la recherche en compréhension liée à la découverte de la relativité et les difficultés conceptuelles liées aux découvertes en mécanique quantique... La troisième crise a été en mathématiques. Elle a été une crise conceptuelle très grande, liée aux notions de rigueur et au moyen d'apporter une preuve mathématique correcte. Par rapport aux anciennes notions sur la rigueur absolue des mathématiques, il est surprenant qu'une telle chose ait pu survenir, et encore plus curieux qu'elle soit survenue dans ces derniers jours où les miracles sont supposés ne pas avoir de place. Et pourtant cela est arrivé.

Remarque : Même après un changement de paradigme, on peut garder des formulations antérieures à titre opératoire. Les règles de calcul de Ptolémée donnent toujours une bonne approximation de la position des planètes. La relativité générale n'a pas fait abandonner l'emploi de la formule de Newton, l'intensité de la force qui s'exerce entre deux corps en raison de leurs masses est calculée avec :

$$f = mm'/r^2,$$

même si l'on peut contester l'emploi de cette expression.

À cette occasion, donnons quelque idée des reproches que l'on peut faire à une formule et plus généralement à l'expression d'un paradigme.

Sur les formules

- L'expression ci-dessus ne contient pas le temps de façon explicite. L'utilisateur peut en déduire, faussement comme on le sait aujourd'hui, que l'interaction à distance se produit instantanément et donc en se propageant avec une vitesse infinie puisqu'il n'est pas dit qu'elle naît au bout d'un certain temps. Newton lui-même avait mis en garde dans ses Principia [\[NEW85\]](#): *«Ce concept de force centripète est mathématique. Je ne juge pas présentement des causes et sièges physiques des forces»*, et encore : *«Lorsqu'il m'arrive de dire que des centres (des corps) attirent ou que des forces y sont appliquées, que le lecteur se garde bien de penser que je définisse, en ces termes, la forme, le mode ou encore la cause ou raison physique d'une action, ou que j'attribue à ces centres (qui sont des points mathématiques) des forces véritables et physiques.»*
- Une formule, même si elle n'induit pas en erreur, ne porte pas nécessairement en elle un témoignage des causes. Celle de Newton contient si peu d'explication sur le phénomène que Richard Feynman [\[FEY70\]](#) a écrit ironiquement : *«Que fait donc la planète ? Est-ce qu'elle regarde le soleil pour voir à quelle distance il est, puis décide de calculer sur sa petite machine interne le carré inverse de la distance, ce qui lui indique de combien elle doit se déplacer ?»*
- Plus encore, la loi de gravitation de Newton peut être écrite de plusieurs façons, équivalentes d'un point de vue mathématique et néanmoins de natures très différentes [\[FEY70\]](#) :

1. l'écriture ci-dessus, la plus connue, simple et d'application aisée;
2. l'écriture de la conservation du moment angulaire;
3. le calcul du potentiel;
4. l'application du principe de moindre action de Fermat et Maupertuis, nommé parfois principe extrémal.

Les deux premières formulations sont de nature locale, les deux dernières sont de nature globale. Ainsi, le même paradigme peut être exprimé par plusieurs formes de lois équivalentes dans le sens où elles produisent les mêmes résultats. Formellement, cette équivalence est celle des expressions mathématiques. Les diverses définitions des [fonctions calculables](#) déjà mentionnées ont un statut voisin.

On peut écrire toute la mécanique sans nulle part faire apparaître la notion de force, en appliquant le point 4 ci-dessus.

En 1894, Hertz pensait déjà que la notion de force n'est pas nécessaire. Peu après, Einstein se demandait si même elle a un sens.

Dans la vision Newtonienne, la force indique à la masse comment s'accélérer ($F=ma$) et la masse indique à la gravité comment exercer une force ($F=GMm/r^2$). L'inertie résiste à l'accélération relativement à l'espace. L'espace agit sur les objets, mais les objets n'agissent pas sur l'espace.

HORS CHAPITRES

Dans la vision Einsteinienne, selon une synthèse fameuse de J. A. Wheeler : L'espace-temps courbe indique aux particules comment se mouvoir et la matière indique à l'espace-temps comment se courber.

Plus généralement un concept physique n'a pas une constitution mathématique absolue qui serait sa vérité profonde, il existe un polymorphisme mathématique. C'est la propriété de ces lois et concepts d'avoir plusieurs mathématisations possibles.

Le mouvement rectiligne uniforme peut être vu :

- géométriquement, espaces égaux parcourus dans des temps égaux par Galilée;
- fonctionnellement, dépendance linéaire entre la distance couverte et le temps;
- analytiquement, vitesse constante ou accélération nulle par Newton.

Un autre exemple est celui de la dynamique du point dans un champ de forces conservatif.

.Elle est formulable par des équations différentielles, Newtonienne;

.par des équations aux dérivées partielles, Hamilton;

.par des principes variationnels, Lagrangienne.

Ces différentes formulations d'une même loi sont mathématiquement équivalentes. Elles ne sont pas physiquement équivalentes, on les distingue par l'histoire du domaine. L'invention d'une nouvelle formulation peut avoir plusieurs causes, besoin de résoudre un problème nouveau ou évolution des mathématiques elles-mêmes.

Sur les paradigmes

À partir du XVII^e siècle et pratiquement jusqu'à nous, la physique est construite sur les [«Principia Mathematica»](#) de Newton.

Les sciences ont connu deux crises d'importances différentes au milieu du XIX^e et du début du XX^e siècle. Elles sont survenues alors que le domaine était très majoritairement :

- . mécaniste à la suite de Galilée et Descartes;
- . fondée sur la physique de Newton;
- . déterministe comme Laplace l'avait affirmé.

Ceci formait un seul paradigme qui a vécu sans vraie contradiction jusque vers 1850. La contrepartie philosophique avait été apportée par Kant. était que l'observation directe par l'homme primait tout. Les lois devaient obligatoirement être déduites des observations directes. À la fin du siècle certains physiciens comme Ernest Mach en viennent à ne se fier qu'à l'observation directe. Ils nient l'existence des atomes car ils ne sont pas directement visibles. Ils sont en cela en accord avec la conception positiviste selon laquelle les concepts majeurs de la physique d'inspiration newtonienne (atomes, masse, espace absolu, temps absolu...) sont des fictions métaphysiques. Ils disent que la science physique s'égaré lorsqu'elle s'interroge sur les «causes» des phénomènes. Elle doit *«se limiter à l'expression des faits observables, sans construire d'hypothèses derrière ces faits, où plus rien n'existe qui puisse être conçu ou prouvé»* E. Mach, la Mécanique, 1883. A. Einstein, physicien classique au début de son activité changea progressivement son point de vue et vint à considérer que les lois peuvent être d'abord imaginées et ensuite vérifiées. La relativité générale et la physique quantique construiront un nouveau paradigme.

Un paradigme est chose très forte. Revenant sur la citation de Laplace, François Lurçat écrit [\[LUR00\]](#): *«On ne saurait accorder trop d'importance à ce texte; des citations répétées ne l'ont pas usé, car l'idéal qu'il exprime a survécu à toutes les objections, à toutes les réfutations rationnelles.»* Le lecteur a remarqué l'emploi du mot «idéal» qui relève plus d'une métaphysique personnelle que d'une pure rationalité.

En mathématiques, l'existence de paradigmes selon Kuhn est controversée. Néanmoins des crises existent en mathématiques comme ailleurs. Dans l'antiquité, il y eut l'incommensurabilité de l'aire et du rayon du cercle (liés par le nombre Pi), ensuite les nombres négatifs, les nombres algébriques solutions d'équations générales de degré supérieur à trois que l'on ne peut pas écrire avec les symboles usuels. [La crise de la première moitié](#) du XX^e siècle a été évoquée plus haut avec la théorie des ensembles, le projet de David Hilbert puis le théorème de Kurt Gödel.

Il arrive que les évolutions ou les transformations de théories ne soient pas complètement maîtrisées. Aujourd'hui, les limites des domaines d'application de la relativité générale macroscopique et de la mécanique quantique microscopique sont encore imprécises. Leurs relations le sont aussi, en attendant une unification éventuelle qui pourrait être apportée par la nouvelle théorie des cordes [\[REC01\]](#). Mais cela stimule plus le physicien que cela ne porte atteinte à la validité de chaque théorie au cœur de son domaine d'application. En mathématiques, il ne semble pas que toutes les conséquences du théorème de Gödel soient encore bien comprises. Des documents et des livres continuent à être publiés sur ce sujet.

Les lois physiques sont justifiées, sauf exception rarissime, à la fois par les observations dont elles doivent être un modèle (phase d'élaboration) et par leurs capacités prédictives (phase d'exploitation). Le fonctionnement des machines est fondé sur les lois de leur domaine. Le schéma *idéal* qui conduit à une machine est alors une chaîne :

observation -> modélisation -> loi -> mise en équations -> résolution -> machine

Cette chaîne produit des résultats pratiques dès qu'elle est établie, mais elle n'est pas indispensable à la construction de machines. Nombreuses sont celles qui ont été construites sur la seule intuition sans théorie explicite ou avec une [théorie très imparfaite](#). Les romains construisaient des moulins à eau sans la théorie de la gravitation. Le fardier de Cugnot est antérieur aux lois de la thermodynamique, Zénobe Gramme a inventé la dynamo avant que Maxwell ait écrit ses équations. Babbage avait conçu sa machine analytique avant que Turing démontre les propriétés de sa machine. Il arrive même que des machines soient construites et fonctionnent alors que leurs capacités sont niées. Une démonstration (fausse bien sûr) de l'impossibilité de voler pour un appareil plus lourd que l'air a été faite dans les premières années du XX^e siècle.

Dans le schéma idéal précédent, une machine est la réalisation concrète d'une solution des équations de la loi en cause, sans liaison nécessaire avec l'usage qui en sera fait. Un moteur électrique peut aussi bien entraîner une machine à laver, un ascenseur ou un aspirateur. Et toujours, le fonctionnement d'une machine dégrade l'énergie qu'elle utilise, selon la deuxième loi de la thermodynamique; l'entropie du système augmente.

Une machine, du simple palan à la fraiseuse, en passant par les moteurs à combustion interne, est fondée sur les lois qui régissent son fonctionnement propre : statique, cinétique, dynamique, élasticité, résistance des matériaux, thermodynamique, etc. Une machine électrique est conçue pour produire ou transformer l'électricité, tous les moteurs électriques sont des solutions, heureuses ou non, des équations de l'électromagnétisme établies par James Clerk Maxwell en 1873. Tant que ces équations seront valides, il ne sera pas envisageable de concevoir une machine électrique en violation de ces équations. La machine construite aura telle caractéristique et telle performance selon la solution choisie.

La machine ainsi construite aura bien sûr des usages divers, mais son utilisation relève toujours de la physique. La finalité, chaleur, mouvement, changement d'état, utilise le produit de la machine : énergie mécanique ou autre accompagnée de l'obligatoire dégradation de l'énergie.

Le changement fondamental des conceptions de la physique fait au début du XX^e siècle par la physique quantique n'entache pas le fonctionnement des machines macroscopiques, même si l'on peut en fabriquer qui sont nouvelles comme les lasers. Pour notre domaine, nous verrons dans la deuxième partie de la conclusion que l'on envisage de fabriquer des nanomachines et des ordinateurs [quantiques](#) dont le fonctionnement physique sera régi par la physique quantique et l'utilisation de ses phénomènes.

Dans le domaine informatique, seules les [machines-langages](#) relèvent d'un schéma voisin du précédent.

observation (de la classe de problèmes à traiter) ->
->**définition d'un langage** (propre à les traiter) ->
->**écriture de l'automate** ->

HORS CHAPITRES

->prise en compte des contraintes techniques ->

->machine réelle.

À la loi et ses équations correspondent ici l'automate et son écriture.

On peut estimer que cette façon de considérer les ordinateurs est générale puisque tout ordinateur est une machine langage de son langage d'assemblage, que celui-ci soit pertinent ou ne le soit pas. Le fait que le cheminement ait été bien ou mal suivi est finalement secondaire. On trouve ici la justification première de l'ordinateur considéré comme l'exécutant de son jeu d'instructions. On a cité de nombreux [enseignements d'architectures](#) qui ont cette ligne directrice.

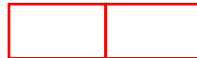
Ce raisonnement séduit par son analogie avec la démarche des physiciens. Le langage et l'automate doivent relever d'une des théories de la calculabilité. À l'intérieur de la théorie de Turing, tout ordinateur universel pourra tout calculer. La machine-langage ne sera pas possible ou impossible, elle sera simplement plus ou moins efficace.

Le schéma général ci-dessus a finalement le même sort que la théorie fondatrice. Sauf à construire une machine par problème à résoudre, ses vertus sont tout de suite épuisées.

Il est utile de s'interroger sur les causes et les conséquences de ces constats, ce que nous tentons dans le fichier suivant.



Conservatoire national des arts et métiers
Architectures des systèmes informatiques
HORS CHAPITRES
Conclusion, deuxième partie
Année 2002-2003



Conservatoire national des arts et métiers
Architectures des systèmes informatiques
HORS CHAPITRES
Conclusion, deuxième partie
Année 2002-2003



Suite N°1...

La vie humaine est limitée ; le savoir est illimité.

Tchouang Tseou Œuvre complète, Gallimard/Unesco, Paris 1969., p. 46

L'idée que le savoir donne des droits [est rétrograde]--on pourrait très bien soutenir que le savoir crée des devoirs.

Jean-Louis Gassée

C 4. SUR LES MACHINES INFORMATIQUES EN PARTICULIER

Venons maintenant à l'informatique parmi les sciences et ensuite aux machines de l'informatique.

Nous venons de voir qu'il existe un schéma général qui mène de l'expression de la loi qui régit un domaine de la science jusqu'aux machines qui en relèvent. Nous commençons donc par une question : l'informatique a-t-elle un statut de science ?

Cette question n'a pas de réponse sûre. Ceux qui tiennent à l'informatique-science sont confortés par la dénomination américaine «computer science». Quant aux autres informaticiens, comme l'écrivait Jean Dieudonné à propos des mathématiciens et de leurs rapports avec le temps et l'infini, 95 % (au moins) d'entre eux ne s'en préoccupent absolument pas.

Plus sérieusement, le modèle déjà présenté de Thomas Kuhn ne paraît pas applicable car il n'y a pas de paradigme affirmé en informatique. Il n'y a pas d'unité de conception entre les architectures de machines, la théorie des langages purement mathématique, les règles relatives aux systèmes d'exploitation, etc. Robert Cori et Jean-Jacques Lévy écrivent dans le cours d'algorithmes et programmation de l'école polytechnique une phrase déjà citée chapitre 2 paragraphe 3 : «*D'abord, il n'existe pas une seule théorie de l'informatique, mais plusieurs théories imbriquées: logique et calculabilité, algorithmique et analyse d'algorithmes, conception et sémantique des langages de programmation, bases de données, principes des systèmes d'exploitation, architectures des ordinateurs et évaluation de leurs performances, réseaux et protocoles, langages formels et compilation, codes et cryptographie, apprentissage et zero-knowledge algorithms, calcul formel, démonstration automatique, conception et vérification de circuits, vérification et validation de programmes, temps réel et logiques temporelles, traitement d'images et vision, synthèse d'image, robotique, ...*»

Cette juxtaposition de théories et pratiques ne fait ni un paradigme ni une science. La cause en est pour nous le fait que l'informatique n'a pas pour objet de modéliser ou d'expliquer des phénomènes.

À l'opposé, Bruno Bachimont [BAC97] écrit : «*L'informatique, à l'instar de la physique, prend comme concept empirique de base un mécanisme algorithmique effectif élémentaire, comme l'application du lambda calcul ou la résolution logique. L'informatique n'est pas une mathématique car elle n'est pas capable de construire a priori son objet dans l'intuition : elle a besoin pour cela d'une existence posée empiriquement pour appliquer les concepts mathématiques. C'est donc bien une physique.*» Cet auteur tient donc pour une informatique de nature physique car opérant sur des objets extérieurs à elle-même. Notre objection est que les dits objets ne relèvent pas de la nature ou mieux ne sont pas des observables extérieurs à l'esprit humain car le «mécanisme algorithmique» est un produit de l'esprit. Nous verrons plus loin qu'il y a au mieux une [relation entre le calcul et la physique](#), elle ne nous paraît pas suffisante pour asseoir un statut.

Ajoutons qu'il n'y a pas non plus une théorie de l'information mais plusieurs théories qui vont des contributions essentielles de Brillouin et de Shannon sur la capacité des canaux de transmission à des théories fortement subjectives qui portent sur les contenus des messages.

Pour une meilleure efficacité, nous proposons de voir et d'examiner les machines informatiques avec le point de vue de la machine double présenté dans le [chapitre 18](#).

La première machine est la machine physique ou support, nous disons la **machine physique**, aujourd'hui réalisée avec des composants électroniques, magnétiques et mécaniques. Elle obéit aux lois physiques.

La seconde machine est la machine hébergée ou superposée, nous disons la **machine symbolique**. Elle est issue de la théorie fondatrice, mathématique, qui n'a aucun lien avec les lois physiques de la première. C'est à cette machine que semble penser B. Bachimont plus haut.

Le lien entre les deux machines est fait par :

- l'inscription des données de la machine symbolique dans la machine physique au moyen d'une mise en correspondance purement conventionnelle entre symboles et phénomènes physiques;
- le fait que des circuits physiques sont fabriqués pour concrétiser les opérations algébriques symboliques sur les phénomènes physiques correspondants.

Les données sont des **objets de la pensée**.
Elles acquièrent une réalité
par leur mise en correspondance avec un **phénomène physique**.

Ce processus de concrétisation est comparable à celui qui fait que l'objet de la pensée «deuxième successeur de un» est représenté par un glyphe, celui du signe que l'on a convenu de lui faire correspondre. Ce type de correspondance a été renouvelé dans l'unification de la [représentation unicode](#) des écritures.

Nous avons constaté que la partie architecturale est constituée pour l'essentiel, au delà des théories multiples de Cori et Lévy, d'une collection d'inventions successives auxquelles la réussite est accordée à trois conditions :

- .condition de principe, ne pas contredire le schéma fondamental;
- .condition de réalisation, être techniquement réalisable;
- .condition de diffusion, apporter un accroissement d'utilité économiquement justifié.

Le concepteur de la machine physique doit réaliser ces inventions, au sens de les faire fonctionner concrètement, avec les outils dont il dispose. Il le fait ou ne le fait pas selon l'état de son art (cf. [microprogrammation](#), [machines Cisc et Risc](#)). À ces inventions s'ajoutent celles qui portent sur la seule machine physique, en électronique ou en magnétisme, inventions et progrès techniques. Elles ne sont pas les moindres comme on le constate aisément par l'observation dite [loi de Moore](#). Les «progrès constants de l'informatique» ont très souvent pour cause les progrès des seules parties de la physique et des techniques correspondantes électronique, magnétisme, mécanique ou optique, appliqués aux machines physiques. Les informaticiens peinent à exploiter utilement ces progrès et nombre d'habits de formalisation taillés par les tenants de l'informatique comme science unitaire ne paraissent être là que pour masquer la pauvreté conceptuelle du domaine.

La machine symbolique manipule des symboles par des opérations réalisées par ses opérateurs. Le fait qu'elle opère sur de l'information n'est déjà pas sur, sait-on ce qu'est l'information? Supposons tout de même qu'il en soit ainsi, elle n'opère a contrario pas sur de la matière ou de l'énergie. Curieusement, cette information est mesurable par des considérations d'entropie sans être de l'ordre de la matière ou de l'énergie. Son aspect énergétique vient de sa correspondante physique (il n'y a pas d'information sans support). Enfin, elle n'est pas utilisée pour faire un travail mécanique mais des opérations. Ceci mérite d'être illustré d'abord a contrario :

Quand on considère la machine proprement symbolique, c'est-à-dire l'architecture informatique et quel que soit le point de vue adopté, la physique n'a qu'une part infime dans les objets manipulés et n'en a pas qui soit connue dans la définition de la machine ni dans la description de son fonctionnement. Le seul lien connu aujourd'hui avec le monde des phénomènes est une relation entre la destruction de l'information et la thermodynamique. Cette destruction obéit à la deuxième loi de la thermodynamique. Détruire l'information augmente l'entropie et on ne peut pas détruire l'entropie. En l'état de la technique électronique, cet effet est noyé dans la dissipation thermique de la machine physique et n'est pas perceptible. Il le deviendra si et quand les effets thermiques des ordinateurs réels seront considérablement réduits. Ce point sera repris dans la troisième partie de la conclusion.

Note : Cette composante symbolique n'existait pas dans les machines analogiques, encore en service à la fin des années 1960. Elles déterminaient des valeurs homothétiques des grandeurs du modèle représenté. Le domaine d'usage des machines analogiques était la résolution de problèmes régis par des équations différentielles. Pour cela on combinait des intégrateurs pour que le temps, variable de base, finisse après intégrations successives par être représentatif de la variable étudiée.

Il n'y a pas en informatique symbolique de matière, de phénomènes ou de grandeurs physiques manipulés pour elles-mêmes.

Les symboles manipulés n'ont ni masse, ni dimension, ni énergie. Nos prédécesseurs avaient raison de nommer machines mathématiques ce que l'on nomme aujourd'hui ordinateurs, non parce qu'elles faisaient des calculs (elles eussent été des machines arithmétiques) mais parce qu'elles traitaient des symboles. On pourrait les nommer machines algébriques en ce qu'elles réalisent des transformations, mais l'impact publicitaire serait mauvais.

La liaison entre les deux machines physique et symbolique est faite par le codage physique. Les données sont codées sous la forme d'états d'un phénomène physique. Le calcul est apporté par la manipulation des données codées. De cette façon, l'information et le calcul sont liés aux processus physiques sous-jacents. Carver Mead nomme son groupe de recherche du Caltech consacré à

la définition des circuits : Groupe de physique du calcul (Physics of computation group).

Bien qu'il n'y ait pas de physique dans la machine symbolique, nous allons raisonner sur les architectures d'une façon qui s'inspire d'elle. Nous avons écrit dans le chapitre 1 que les réalisations que l'on observe : schémas, modèles de mémoire, parallélismes divers, ne sont pas fournies par le monde physique. Elles ne sont donc pas des observables dans le sens de phénomènes fournis par la nature. Par ailleurs, la machine symbolique n'est pas déductible d'une théorie au delà de son principe. Les réalisations sont ainsi sans statut propre, nous les considérerons comme des observables par défaut car :

1. elles ont une réalité objective;
2. elles sont pas filles d'une théorie par déduction;
3. sauf à n'en rien dire, on doit s'en préoccuper.

En leur donnant le statut au moins provisoire d'observables, nous procédons comme le fait le physicien qui, ayant une loi dans un domaine de validité éprouvé, se demande si cette loi ne serait pas applicable ailleurs. Dans notre cas, une étude systématique pourrait conduire à construire une théorie de ces réalisations, théorie probablement inhabituelle et ne correspondant pas aux canons usuels. Ses prémisses figurent dans le chapitre 18.

CE QUE L'ON OBSERVE DANS LA MACHINE SYMBOLIQUE ET CE QUE L'ON N'OBSERVE PAS

Les invariants

Le plus grand nombre des lois de la physique sont des lois de conservation, conservation de la masse, de l'énergie, de moments angulaires, etc. Elles constituent la classe très générale des lois génératrices d'invariants. Elles sont écrites en termes d'égalités ou de bilans. Un exemple bien connu est celui de la loi d'Ohm :

$$V - RI = 0$$

Tenant R constant, toute modification de V fixe une valeur de I et réciproquement. C'est volontairement que nous n'avons pas écrit $V = RI$ qui n'a pas la forme d'un bilan.

On ne trouve pas d'invariants en informatique. En quantité, les données sont multipliables ad libitum par des programmes simples. On peut remplir la mémoire et des rames de papier par des listes de nombres entiers ou de suites de symboles. On ne crée pas d'information au sens de la réduction d'incertitude mais une accumulation de données avec ou sans signification. On peut aussi calculer des millions de décimales de π ou des nombres entiers successifs et les écrire.

Réciproquement un programme court et sans complexité peut effacer une quantité considérable de données. Un programme de formatage physique du support détruit toute information.

Il n'y a pas de loi d'invariance.

La localité

La physique macroscopique a des propriétés locales, alors que ce n'est pas le cas en physique quantique, cf. les expériences d'Alain Aspect [\[ESP79\]](#). Un événement influence fortement ses voisins, plus faiblement des éléments éloignés. Cette influence décroît avec la distance. La loi de Newton est en $1/r^2$ même si les actions ne sont pas instantanées. Les réactions chimiques se

font par contact. Les propagations se font de proche en proche, toujours avec atténuation. Dans les machines analogiques, il fallait veiller soigneusement à conserver la grandeur d'une tension.

Il n'y a pas une telle localité en informatique. En mémoire centrale, un mot est atteint dans un temps indépendant de sa position dans la machine symbolique et peu s'en faut dans la machine physique. Un processus logé en mémoire modifiera indifféremment des données quelle que soit leur distance. Plus même, la condition d'efficacité du processus est qu'il opère sur les données quelle que soit leur localisation en mémoire, sur disque ou même très loin via un réseau. Dans les communications, la distance intervient mais on s'intéresse plus au débit, que de trop nombreux informaticiens nomment vitesse, qu'au délai d'acheminement qui est une contrainte de la machine physique, inexistant pour la machine symbolique. La latence est un inconvénient que l'on cherche à masquer et non un phénomène propre au fonctionnement symbolique. Le délai de propagation sur une ligne est proportionnel à la racine carrée du produit LC, où L et C sont les valeurs d'impédance et de capacité par unité de longueur. Les tensions se propagent sur les bus à des célérités qui sont au mieux de 15 à 20 cm/ns soit entre 150000 et 200000 km/s. Ceci n'apparaît en aucune façon dans l'écriture des instructions.

Il n'y a pas de localité.

L'énergie

Les opérations physiques ont toujours une composante énergétique, par exemple un échange d'énergie potentielle due à un champ ou un gradient avec d'autres énergies : énergie gravifique et énergie cinétique pour une chute d'eau, différence de potentiel et courant électrique, échange d'énergie et de masse dans une réaction nucléaire, etc. Ces échanges sont régis par les lois d'échange ou substitution dont les modèles sont des équations de conservation.

Le support de la machine informatique est de nature mécanique si l'on conserve l'ancienne acception du terme. Il comporte donc des échanges énergétiques. Ceux-ci sont propres aux solutions techniques choisies. On dit que les brevets de Gene Amdahl sont pour la plupart relatifs à la thermodynamique.

La machine symbolique n'est pas énergétique. Les échanges d'énergie ne concernent que la machine électronique. Bien qu'inévitables, ils ne doivent pas influencer la machine symbolique. Il n'y a donc pas de lien nécessaire entre une opération informatique de manipulation de symboles et l'énergie qui lui est nécessaire, même si l'on ne peut pas envisager une machine à consommation nulle. Au cours du temps, on a construit des machines toujours moins consommatrices d'énergie pour faire les mêmes opérations et les faire considérablement plus vite, de la machine à relais aux circuits BiCMOS, en passant par les lampes, les transistors, les variantes PMOS, NMOS, etc.

Peut être existe-t-il une exception à cela qui porte sur l'effacement des informations. Cet effet n'est pas perceptible dans les machines actuelles, on lira pour cela les considérations relatives à [l'entropie du système](#).

Il n'y a pas d'énergie.

Le bruit

Le fonctionnement de la machine physique est affecté de bruits. Ces bruits sont souvent gênants, parfois utiles, ils sont maîtrisés ou ne le sont pas, mais ils sont toujours présents et acceptés comme faisant partie du fonctionnement. Ce sont les frottements mécaniques, le bruit de grenaille, des bruits quantiques, etc.

La machine symbolique n'a pas de bruit. Le bruit existe dans la machine physique, mais les symboles manipulés ne doivent pas en être affectés et des précautions sont prises pour cela. Il y a des fautes et des erreurs dans la machine symbolique mais pas de bruit.

Il n'y a pas de bruit.

Le nombre d'états

Les machines physiques à fonctionnement discontinu ont un petit nombre d'états distincts. Il en est de même dans les machines à fonctionnement apparemment continu où les états sont discernables par la précision des instruments de mesure, ou par la quantification compatible avec le bruit de fonctionnement.

Une machine informatique a un nombre d'états distincts qui relève de la combinatoire. Un banc de mémoire d'un Mégaoctets a $256^{(1024 \times 1024)}$ états distincts soit plus de $256^{1000000}$, et ils peuvent être atteints de façon uniforme. Par comparaison, le nombre d'atomes estimé de l'univers est inférieur à 10^{80} .

Le nombre d'états distincts relève de la combinatoire.

DERNIER RETOUR SUR LES ARCHITECTURES

L'un des défauts de notre époque moderne est de penser que le présent est la somme et le couronnement de tout le passé.

Nous avons, dans ces notes de cours et oralement, écrit et dit plusieurs fois qu'un ordinateur a des architectures différentes selon le point de vue que l'on prend. Il n'en reste pas moins que la machine matérielle, machine physique support a une architecture précise, faite pour l'essentiel de bascules, de capacités, de chemins et de circuits.

Cette architecture est différente d'une machine à l'autre. Ses évolutions sont diverses. Pourquoi donc une seule architecture ne va-t-elle pas d'elle-même ? La réponse, on l'a établi, vient de ce que le schéma de Neumann est indifférent aux modalités de réalisation et aux exigences des utilisateurs en matière de performances, vis-à-vis desquelles les architectures ne sont pas neutres.

Pourquoi alors ne pas s'être accordés sur une seule architecture ?

À cette première question, nous avons donné un élément de réponse qui tient à la différence entre calcul arithmétique et manipulation de symboles.

Pourquoi telle architecture efficace à un moment ne l'est plus à un autre puisqu'on constate son

remplacement?

À problème compliqué et impossible à poser mathématiquement, il est rare d'avoir une réponse simple. Nous retiendrons une première réponse de principe qui n'apporte pas grand chose à la compréhension, et trois réponses plus détaillées.

- *La réponse de principe* est que la capacité technique augmente et que des idées nouvelles apparaissent, même rarement.

Les trois réponses plus détaillées ont besoin de plus d'explications :

- *La première* est la plus propre nous flatter, un homme a une idée nouvelle. Le fait est rare, on l'a dit. Il arrive à tout moment, souvent sous l'empire de la nécessité, parfois en conséquence d'une obsession, peut-être par pur hasard. Il est donc relativement indépendant du progrès des machines support. Un exemple significatif est le registre d'index en 1951. Remarquons à nouveau qu'en matière architecturale, la production de nouveautés date des années 1950 et 1960.
- *La seconde réponse* prend en compte le progrès technique. Il se manifeste par une augmentation des capacités de mémoire et des débits de traitement. L'ampleur de ces évolutions peut être faible ou grande. Si elles sont de faible ampleur, on n'observe pas de changement drastique dans les architectures. La pesanteur des habitudes et la nécessaire rentabilisation des investissements s'y opposent. Quand le progrès fait changer les paramètres d'un ordre de grandeur environ, alors il peut y avoir changement, évolution importante ou substitution. Reprenant l'évolution des microprocesseurs d'Intel (cf. annexe [asi0008](#)), on notera que pendant de nombreuses années, le processeur suivant surclassait son prédécesseur d'un facteur dix, plus ou moins. L'architecture de principe y a été modifiée fortement deux fois et des modifications substantielles ont été apportées plusieurs fois. Parmi ces évolutions nous noterons l'ajout de coprocesseurs pour le calcul en virgule flottante, l'ajout d'unités de gestion de mémoire. Elles proviennent bien de la capacité que l'on a d'ajouter des circuits spécialisés, à coût presque constant.
- *La troisième réponse* fait intervenir un facteur plus subtil, dû à Jean-Pierre Meinadier. Le progrès technique de volume ou de débit n'intervient pas par lui-même mais par les différences qu'il induit entre les performances des composants. Pendant que la fréquence d'horloge des processeurs passait de 10 à 500 MHz (facteur 50), la vitesse de rotation des disques passait de 3600 à 7200 tours par minute (facteur 2) et les temps d'accès aux mémoires dynamiques passaient de 120 à 60 ns (facteur 2). Ces changements, par le fait qu'ils n'étaient pas uniformes, ont induit le surgissement des caches de toute nature, la création de bus multiples spécialisés plus récemment et après la mise en œuvre de la microprogrammation dans les années 60 rendue possible par les mémoires électroniques, l'apparition des machines RISC. Les évolutions architecturales tiennent plus à des différences de progrès entre éléments qu'à une évolution générale indifférenciée.

Sur les architectures que l'on observe.

Une conclusion d'ordre pratique peut être tirée des nombreuses observations déjà faites et des trois réponses précédentes. Les tendances que l'on observe peuvent rapidement être transformées par une percée technique sur un point particulier. Donnons deux exemples parfaitement hypothétiques et pourtant réalistes.

- Supposons que la technique d'enregistrement magnétique continue à progresser au rythme actuel et que la gravure des circuits passe de 0,18 nm à 0,05 nm. La taille de la mémoire centrale à coût pratiquement constant passera de 256 Mo à plusieurs gigaoctets. Que deviendra alors la mémoire virtuelle si la mémoire câblée a la capacité de charger la quasi

totalité du disque ?

- Dans les mêmes conditions les mémoires SRAM auraient des capacités de plusieurs dizaines de mégaoctets. Garderions nous plusieurs étages de caches ou un seul suffirait-il ?

Et fin du commencement

Les premiers mots du premier chapitre étaient : «L'architecture est le produit de l'art de l'architecte». En matière d'art, Ernst Gombrich a écrit : «Il n'y a pas d'Art, il n'y a que des artistes.» Nous savons depuis le début du second chapitre pourquoi il n'y a pas **une** informatique. Dans le premier chapitre nous avons vu ce que les grands acteurs ont apporté. Peut être n'y a-t-il pas d'informatique mais seulement des informaticiens.

Le même Ernst Gombrich explique que le progrès n'existe pas dans l'art. Selon lui, les civilisations et les époques changent ce qu'il nomme les situations. L'artiste égyptien ne représentait pas dans une sépulture ce qu'il voyait mais les objets nécessaires à la vie. Le peintre roman écrivait des symboles pour enseigner le peuple illettré. L'auteur de fresques byzantines voulait imprégner le fidèle d'une atmosphère religieuse précise.

La course à la puissance ou prétendue telle obnubile l'informaticien. Et pourtant il n'est que spectateur et rarement acteur devant ce que produisent les électroniciens. Ces capacités croissantes ont un sens et se dirigent vers toujours plus. Elles sont portées par un progrès technique.

Mais l'informatique a-t-elle une direction ou son progrès a-t-il un sens ?

Les premiers de nos prédécesseurs voulaient faire des calculs complexes ou à tout le moins nombreux. Ceux qui ont encore ce but sont devenus une toute petite minorité, météorologues, physiciens, chimistes, etc. Pour le plus grand nombre, le calcul n'est plus la situation même si les processeurs traitent toujours plus vite plus de symboles.

Ensuite sont venus les hommes de la gestion. Leur situation était de trier, d'ordonner, de conserver des fiches et de produire des états. Ils sont survenus quand le progrès technique leur a apporté des entrées et sorties simultanées ou presque, des supports de masse, et une fiabilité acceptable qui ne menait plus au désespoir.

Une autre situation est celle de la production des biens matériels pour laquelle on a fait des bus et des réseaux de terrain, des moniteurs de temps réel et des automates programmables.

Bien plus tard sont venus les traitements des sons et des images.

Aujourd'hui arrive la réalité virtuelle.

Il serait utile de vérifier que les architectures, non pas en elles-mêmes mais dans leur existence à un moment donné, sont plus et mieux explicables par ces situations que par la puissance, non mesurable on l'a vu et bien mal nommée de la machine disponible. A-t-on seulement les architectures que l'on demande? Les concepteurs et fournisseurs proposent les architectures dont ils connaissent ou subodorent la demande. La considération des besoins propres aux situations apporte une signification.

- Prévoir le futur en terme de Mips n'est pas vraiment difficile. Les laboratoires détiennent les matériaux des années à venir et ce qui sera possible en GHz et en micromètres.
- Prévoir le futur en termes des situations est beaucoup plus difficile, alors que c'est là que peut résider une clairvoyance utile. Que sera la demande ? Les clients voudront-ils

plus de puissance pure ? Peut-être pas, la puissance est utilisée, elle n'est pas cherchée pour elle même.

L'explication par le «pour quoi faire» n'apporte pas de garantie. Il suffit de se rappeler ce qu'il en fut pour quelques produits aujourd'hui bien connus :

- le téléphone à son début pour écouter les pièces de théâtre chez soi selon Graham Bell;
- le téléphone plus tard comme outil professionnel et surtout pas à des fins personnelles ou privées selon la Direction générale des télécommunications en France dans les années 1950 et 1960;
- la télévision comme outil pour s'instruire selon nombre de journalistes et d'enseignants;
- le minitel pour consulter des banques de données selon ses promoteurs, etc.

La prévision des situations est tout sauf facile. Au moins peut-on prendre claire conscience du passé, faute de connaître l'avenir.

Le passé contient l'histoire des idées encore plus que celle des machines diverses du premier chapitre. Ce n'est pas une suite d'événements mais un déroulement. Ces idées ont existé, existent encore ou ont été abandonnées. Nous ne sommes pas les premiers à réfléchir. Les idées et d'autres notions peuvent être regroupées sous le mot plus général de philosophie, au singulier ou au pluriel.

Trop nombreux sont ceux qui ont une attitude de refus quand ils lisent ou entendent ce mot. Cette attitude est, mutatis mutandis, de même nature que celles de beaucoup de scientifiques et surtout des matérialistes devant le mot métaphysique, par ignorance, par conditionnement ou par mode. Les causes ou les raisons de ces refus ne sont pas notre propos, chacun a ses amours et ses préférences, ses aversions et ses phobies. Que Cicéron ait écrit «Rien ne peut être dit de si absurde qui ne soit dit par quelque philosophe» ne doit pas faire rejeter toute réflexion. Que la philosophie régnante soit aujourd'hui l'idéalisme depuis Descartes, Spinoza, Kant, Fichte, Heidegger et d'autres ne fait pas oublier la racine aristotélicienne réaliste comme l'ont été Bergson et d'autres il y a peu.

Nous avons mentionné Aristote dans le premier chapitre pour ses apports en matière de raisonnement et plus spécialement pour l'établissement du syllogisme. La logique comme discipline est aujourd'hui fréquentée autant par les philosophes que par les mathématiciens et les informaticiens, même s'ils n'en parlent ni ne l'utilisent pas de la même façon. Son objet premier est l'étude des lois du raisonnement correct. Elle mène à distinguer la matière de la forme dans un raisonnement, c'est-à-dire ce sur quoi il porte, distingué de son expression. Le champ de la logique est la vérité, en vertu de la forme seulement; on peut dire en vertu de la syntaxe de son écriture.

C'est au XIX^e siècle que la logique est devenue mathématique par les travaux de George Boole et de Gottlob Frege notamment, alors que la logique de Port-Royal se préoccupait encore de sa matière et de son bon usage. Le titre de l'ouvrage principal de Boole est : "An investigation of the laws of thought on which are founded the mathematical theories of logic and probabilities"; une étude des lois sur lesquelles sont fondées les théories de la logique et des probabilités. L'objectif était alors de substituer la rigueur et l'exactitude du calcul mathématique à l'imprécision des contenus et parmi eux, ceux des concepts philosophiques.

Précisons ces points : En logique mathématique, les raisonnements ou propositions sont vrais par le seul fait de leur forme, indépendamment de leur matière qui est pourtant ce dont il est question. La validité matérielle, issue du contenu n'est plus considérée car la valeur, vérité ou fausseté, change avec le contenu : le soleil est dans le ciel n'a pas de rapport de vérité avec la terre est brune. On veut que la matière puisse varier indéfiniment sans que la valeur en soit affectée : si A implique B et B implique C, alors A implique C quels que soient A, B et C.

Dans un second temps, la logique mathématique a été utilisée à toute autre chose. Étant la science du raisonnement, elle est aussi celle de la démonstration, outil essentiel des mathématiques. La logique va être utilisée pour fonder les mathématiques en UNE mathématique parfaitement établie sur le nombre le plus petit possible de présupposés. On a présenté les projets de D. Hilbert et des travaux de K. Gödel et A. Turing dans le premier chapitre.

Le syllogisme fondement du raisonnement devient celui de la démonstration. Il est utilisé en l'intelligence artificielle. Alan Turing était persuadé qu'on en viendrait à donner une intelligence aux machines.

La philosophie désigne habituellement ses orientations à l'aide du suffixe -isme qui qualifie, le plus souvent, une école de pensée ou tout au moins un point de vue bien constitué, Platonisme, Aristotélisme, réalisme, idéalisme. En mathématiques on connaît de même les formalisme, intuitionnisme, constructivisme, logicisme, conventionnalisme. L'informatique est trop jeune pour que ces mots en isme soient éclos. Ne désespérons pas, on aurait pu parler de centralisme et d'individualisme dans les disputes entre informatique centralisée et informatique individuelle.

Les questions sur lesquelles certains de nos contemporains s'interrogent sont notamment :

- **L'esprit est-il une machine ?** Une machine peut-elle penser? Ces questions ne sont plus seulement philosophiques. La neurologie, une meilleure connaissance du cerveau humain et les sciences dites cognitives y contribuent. Sur ce sujet, Bruno Bachimont écrit [\[BAC97\]](#)

"Les sciences cognitives ont pour but d'appliquer à l'esprit les méthodes d'investigation des sciences de la nature : il s'agit de naturaliser l'esprit." Et aussi : "Les sciences cognitives sont issues d'une tradition philosophique et scientifique permettant de reposer à nouveaux frais la question de l'esprit. La tradition logique et formaliste aboutit au concept d'ordinateur dans lequel le cognitivisme computationnel voit l'essence même de l'esprit, comme raison calculante. La tradition biologique et évolutionniste voit dans l'esprit une propriété du vivant fruit de l'évolution dont il faut définir les caractéristiques depuis une théorie du vivant. La tradition physique et mathématique reconnaît l'esprit dans la description qualitative d'un système dynamique particulier, le cerveau. Ces théories, riches et complexes, ne vont pas sans problèmes profonds, renvoyant à des difficultés soulevées depuis parfois fort longtemps en philosophie. Il faut donc interroger la philosophie pas seulement pour comprendre l'origine des sciences cognitives, mais pour problématiser leur échec et formuler leur critique."

Le lecteur observera que le "cognitivisme computationnel", comprendre "par le calcul", bien qu'issu d'une tradition honorable est fortement réducteur et se trouve loin de couvrir la totalité des activités de l'esprit : l'acquisition de connaissances nouvelles d'origine extérieure n'y figure pas, l'induction, forme d'imagination, non plus. La logique et ses avatars ont toujours été un moyen, un outil ou un art de raisonner, mais non de ressentir ni de comprendre. On peut apprendre sans comprendre. La théorie de Proust sur le style d'écriture est que le style est une vision qui n'est pas analysable par l'intelligence.

- **L'évolution de la pensée** ou encore l'émergence de formes d'idées nouvelles. Des auteurs ont identifié quatre étapes ou niveaux de civilisations : la parole, l'écriture et la lecture, l'imprimerie, le PC avec internet et le web. L'écriture a apporté des notions nouvelles par la mise en situation géométrique de la pensée, auparavant fortement temporelle. Les listes, les tableaux en font partie. Nous disposons aujourd'hui d'un nouveau support, actif ou dynamique, l'ordinateur et un réseau quasiment illimité. Il est possible que des effets comparables apparaissent. Les premiers pourraient être liés à la capacité de créer des objets animés et non plus des textes ou des images sur écran ou sur papier.

- **La relation au réel.** Les philosophes connaissent bien les différences de conception qui

séparent les idéalistes des réalistes. Pour les premiers, la réalité concrète au moins n'a que peu de sens, au plus n'existe pas; seul notre esprit existe et pense le monde. Pour les seconds, le concret existe hors de notre esprit qui s'en forme image par nos sens. L'apparition de la réalité virtuelle peut changer ces façons de comprendre ou d'interpréter la relation entre l'esprit et le réel.

- **L'universalité.** Le lecteur a remarqué l'emploi de l'adjectif *universel* pour la machine de Turing et les ordinateurs. Cet adjectif est aussi utilisé en physique pour qualifier les lois. Cette façon de parler est habituelle chez les réductionnistes qui estiment que tout peut être décrit par des lois et simulé par calcul. Nous allons voir que la signification d'universel n'est pas la même dans ces deux usages.
 - En physique, l'universalité de la loi est limitée à un ensemble bien précis d'observables. Pour eux, elle doit être valide toujours et partout, dans toutes les conditions d'observation. Ceci est indispensable pour distinguer une loi d'une observation empirique, mais ce n'est pas vérifiable.
 - En matière de calcul, nous n'avons pas affaire à des observables au sens de la physique, l'objet est un jeu de récritures de chaînes de symboles. Si l'on examine la chose de près, on s'aperçoit que la machine de Turing définit la classe des fonctions calculables en son sens. La machine de Turing est donc universelle au sens de la classe qu'elle a créé. Pour sortir de ce cercle, il faut admettre les [thèses](#) de Church et Turing que nous rappelons :
 - **thèse de Church:** Toute procédure constructive, finie, de calcul peut être réalisée par une machine de Turing;
 - **thèse de Turing:** Aucun système formel dont le calcul est défini comme une séquence finie d'étapes n'est plus puissant que le modèle de calcul de la machine de Turing.
- **Les limites.** Quelles sont donc les limites connues de ces universalités limitées? En physique la plus connue est celle des relations d'incertitude. En matière de calcul, c'est le théorème de non consistance; des fonctions existent qui ne sont pas calculables.

En résumé :

Question 1. Est-ce qu'un ordinateur peut penser ?

Beaucoup de formes de comportement intelligent semblent descriptibles comme des fonctions réglées produisant des sorties à partir d'entrées : planification, résolution de problèmes, jeu d'échecs, production de langage, etc. Une machine de Turing peut, en principe, exécuter les programmes qui réalisent de telles fonctions. Est-ce qu'une telle machine est alors intelligente ?

Comment savoir si une machine pense ?

Que voudrait dire «avoir conscience de lui-même» pour ordinateur ?

etc.

Question 2. Sommes nous des ordinateurs ?

Est-ce que notre cerveau est une réalisation d'ordinateur virtuel tout comme une machine de Turing réalise un ordinateur virtuel ?

Est-ce que la pensée est réglée par la logique ?

Est-ce que le cerveau opère sur des symboles ?

Est-ce que tous les phénomènes mentaux peuvent être décrits en termes de calculs ?

etc.

Tout ceci en tenant compte du fait que, quoi qu'il en soit, une machine n'est pas un organisme vivant.



Conservatoire national des arts et métiers
Architectures des systèmes informatiques
HORS CHAPITRES
Conclusion, deuxième partie
Année 2002-2003



Conservatoire national des arts et métiers
Architectures des systèmes informatiques
HORS CHAPITRES
Conclusion, troisième partie

Année 2002-2003

--	--

--

*Yeou, savez vous ce que c'est que la science ?
Savoir que l'on sait ce que l'on sait et savoir que l'on ne sait pas ce que l'on ne sait pas.
Confucius, Lun-yu ou les entretiens philosophiques.*

C 5. SUR LES FUTURS POSSIBLES, MÉTHODE DE PRÉSENTATION

Nous adoptons pour cette partie l'ordre qui va dans du proche au lointain dans le temps :

1. un bref résumé des tendances récentes;
2. des évolutions déjà acquises par l'existence de prototypes probants ou de maquettes, jusqu'à 2005 plus ou moins;
3. des modifications moins certaines par le fait qu'elles contiennent des facteurs de risque non maîtrisés, jusqu'à 2015 plus ou moins;
4. des changements simplement esquissés, même si des travaux leur sont consacrés.

Dans chacun de ces groupes, nous distinguerons autant que possible ce qui a trait à la machine support et à la machine symbolique.

Le futur ne se limite pas à des évolutions scientifiques et techniques. Le conditionnement des produits, leur usage comptent beaucoup. Cette affirmation peut être confortée par plusieurs exemples.

Pour le conditionnement : dans plusieurs domaines la vente à assembler (le kit) apporte diffusion et moindre coût. Les premiers micro ordinateurs étaient les un ouverts, les autres fermés. Le succès récent de l'Imac d'Apple doit beaucoup à sa forme.

Pour l'usage : il arrive qu'il n'est pas celui que l'on envisageait au départ. Le téléphone était destiné à écouter chez soi des pièces de théâtre, le minitel doit plus aux messageries roses qu'aux bases de données pour lequel il était fait. Internet est construit sans innovation technique.

Le futur sera fait de conditionnements nouveaux, de nouvelles idées d'utilisation. Les prototypes actuels sont divers : la face interne d'un verre de lunettes comme écran et l'ordinateur dans la poche; l'ordinateur de 5 cm d'épaisseur derrière un écran plat.

L'auteur n'est pas devin et n'a pas la capacité d'inventorier tous les projets existant de par le monde. Le lecteur trouvera dans la suite ce qui lui paraît être porteur d'avenir, avec les restrictions normales sur le devenir d'évolutions possibles mais non nécessaires.

Notion d'innovation contrainte.

Les évolutions présentées comme probables ou seulement possibles doivent être considérées, non pas comme des innovations libres mais comme des **innovations contraintes**. Il y a trois contraintes principales.

- Le maintien de la **compatibilité avec les jeux d'instructions** des machines précédentes quel que soit le moyen employé pour cela et il est parfois lourd. Intel a longtemps conservé le jeu x86 au prix d'une grande complexité de ses microprocesseurs. Plus récemment il a utilisé les micro opérations. Les PowerPC contiennent une couche d'émulation pour le même jeu x86. La nouvelle architecture de processeur d'Intel, nommée IA64, a un mode compatible IA32 qui lui-même exécute le x86.
- Le maintien de la **compatibilité avec le système d'exploitation** des machines précédentes.
- Le maintien du **modèle d'exécution séquentielle**, moins pour obéir à J. de Neumann que pour conserver la notion d'état de la machine quand survient une interruption et pouvoir ainsi reprendre l'exécution ultérieurement.

Un programme est :

de plus en plus la liste de ce qu'**il faut faire**;

de moins en moins la spécification de **comment le faire**.

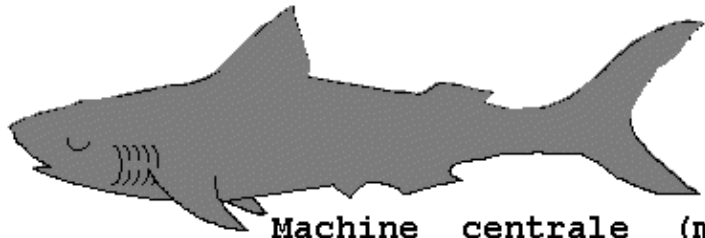
C 6. LES FAITS IMPORTANTS DES DERNIÈRES ANNÉES

David Patterson a construit l'image parlante de la chaîne alimentaire où les ordinateurs sont représentés par des poissons, la question est : «qui va manger l'autre?» (<http://www.cs.berkeley.edu/~pattarn/252S98/index.html>)



Les gros poissons mangent les petits

1988



Machine centrale (mainframe)



Superordinateur



Mini-superordinateur



Mini-ordinateur



Station de travail

PC



Ordinateur massivement parallèle

En 1988, ils sont encore

tous là.



Ordinateur
massivement parallèle



Mini-
superordinateur Mini-
ordinateur



Machine centrale (mainframe)



Serveur

Station
de travail PC



Superordinateur

En 1998, déjà deux

1998

Qui va manger qui ?

disparus !

Les deux derniers poissons qui ont été mangés sont :

- les mini-superordinateurs (1963-1996) quand Cray Research a été acheté par SGI;
- les mini-ordinateurs (1965-1998) quand DEC a été acheté par Compaq.

Un autre fait notable :

Le laboratoire national américain Sandia et Intel ont construit en 1996 un superordinateur qui a atteint 1 000 milliards d'opérations flottantes par seconde (banc d'essai [Linpack](#)) par seconde, retrouvant la masse d'une machine à lampes : 44 tonnes, mais avec 573 Go de mémoire et 2,25 milliards d'octets sur disque. C'est une machine massivement parallèle contenant 7 264 PentiumPro sur un total prévu de 9 200.

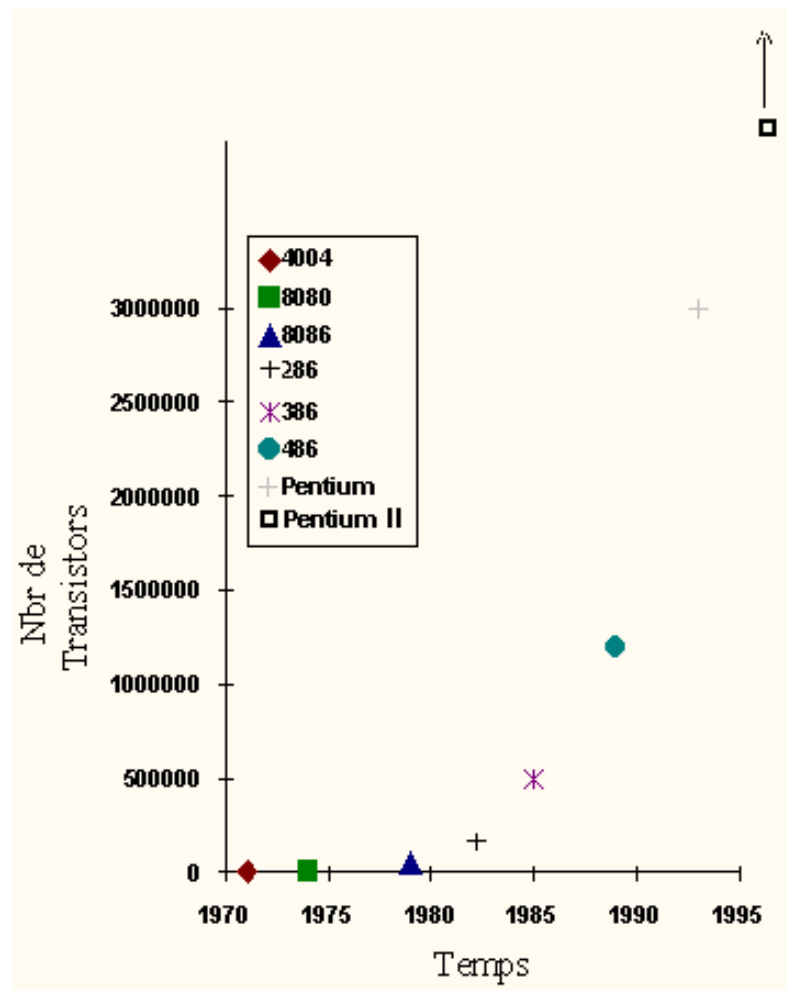
C6.1. La machine physique pendant les dernières années

C6.1.1. En matière électronique

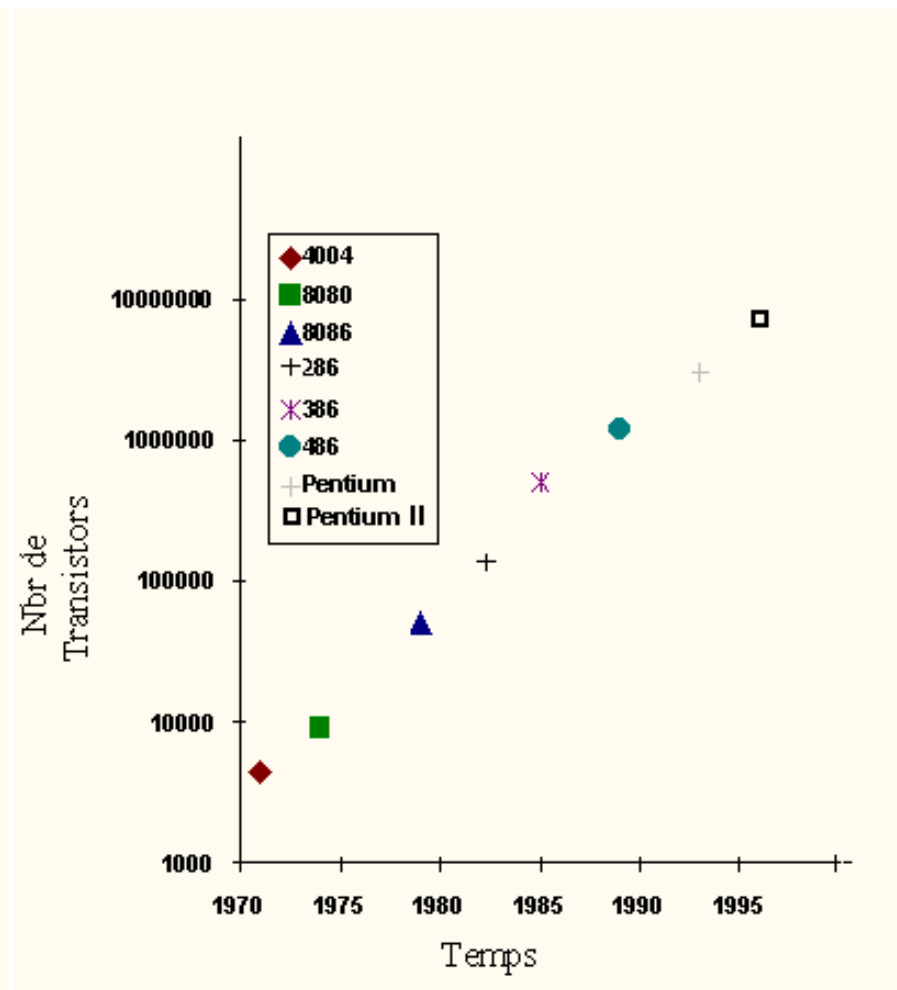
La loi de Moore déjà citée observe que le nombre de transistors dans une puce suit une progression temporelle géométrique à coût constant depuis plus de 25 ans. Le doublement est fait en 18 mois plus ou moins. On trouvera ailleurs les [nombres de](#)

[transistors](#) des microprocesseurs Intel ainsi qu'un tableau récapitulatif des [fréquences des microprocesseurs](#). Les deux figures ci-dessous sont à gauche en échelle linéaire, on y perçoit assez bien la croissance exponentielle passée. La figure de droite en échelle logarithmique établit encore mieux le caractère exponentiel pendant 25 ans par le quasi alignement des points.

Échelle linéaire



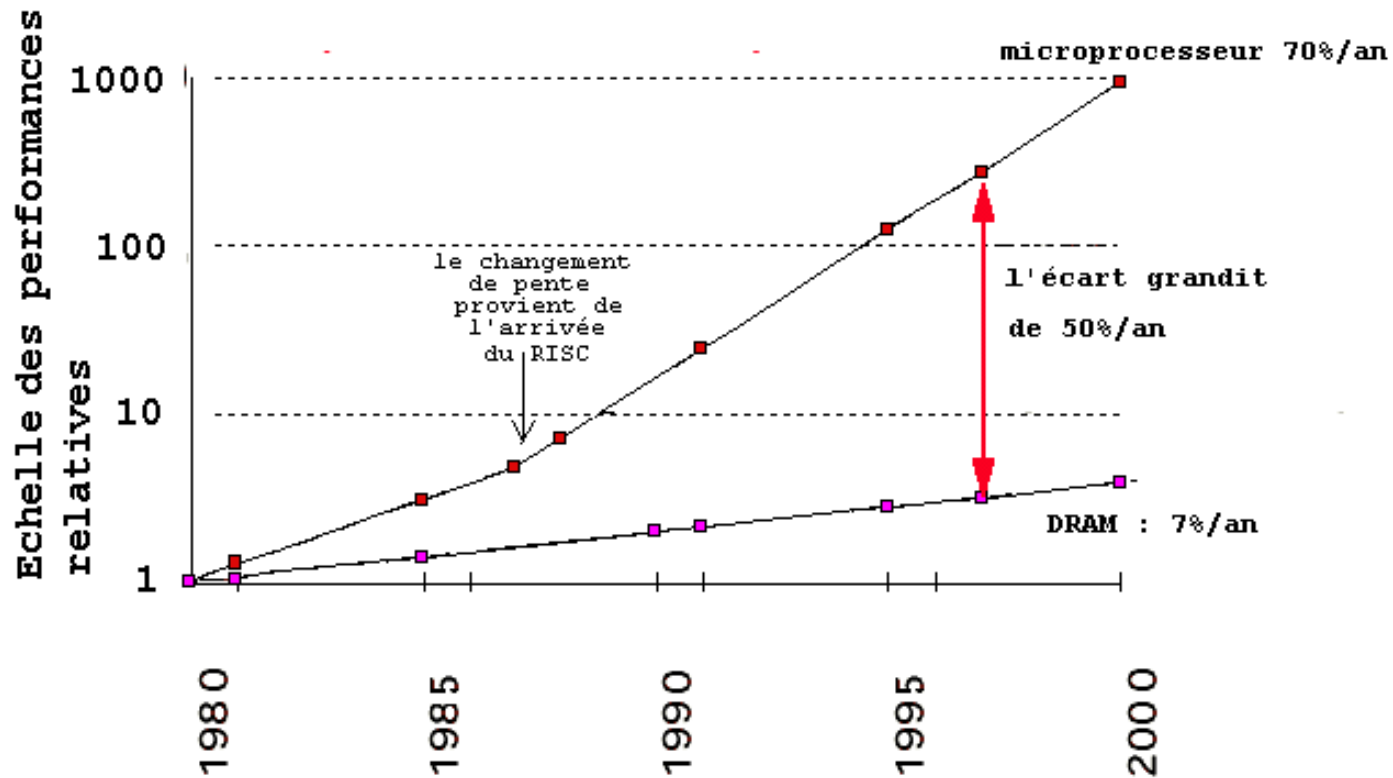
Échelle logarithmique



D'autres auteurs le disent de façon différente. Michel Dagenais, de l'institut polytechnique de Montréal au Canada écrit :

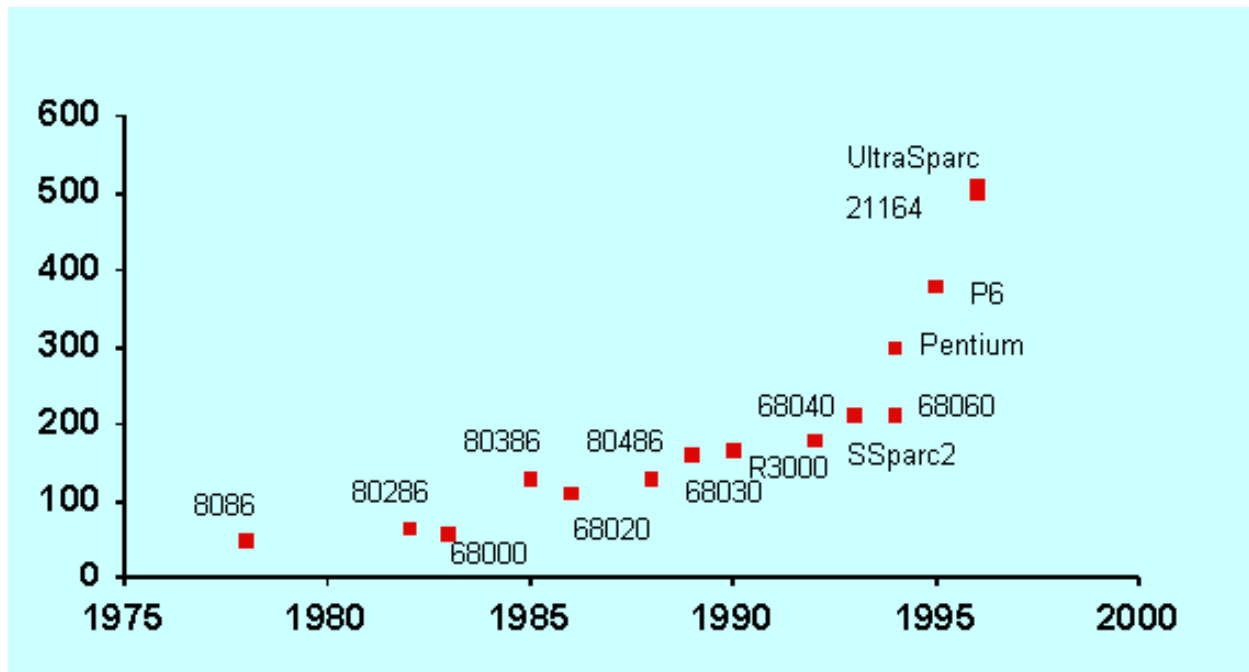
les circuits logiques sont 25% plus rapides et denses chaque année;
 la mémoire vive est 60% plus dense chaque année, 30% plus rapide en 10 ans;
 les disques sont 25% plus denses chaque année, 30% plus rapides en 10 ans.

Un résumé parmi d'autres de cette évolution est fait dans la figure suivante :



Cet écart considérable et croissant est la justification des caches.

Dans le même temps le nombre de pattes des puces processeurs est passé de 64 en 1978 à plus de 500 comme dans les DEC Alpha 21164 et UltraSparc et 603 sur un Pentium 4 Xéon d'Intel. Pendant ce temps, les bus de données sont passés de 8 bits à 128 bits (Powerpc G4) et les bus d'adresses de 16 bits à 40 bits (Alpha 21164). En plus, l'augmentation du nombre de pattes a été nécessaire pour l'alimentation en puissance et la mise à la masse des grandes surfaces de puces.



C6.1.2 En matière de supports magnétiques.

On a observé plusieurs évolutions qui ont affecté les [caractéristiques de disques](#) dans les années passées.

La vitesse de rotation est passée de 3600 t/mn à 10000 t/mn avec une réalisation à 15000 t/mn. On n'attend plus d'augmentation significative.

Les [bus interfaces](#) ont beaucoup évolué.

Un cache de 1 Mo est devenu banal.

La densité d'enregistrement magnétique est passée de 10 Mbits/pouce carré en 1980 à 1 Gbits/pouce carré à la fin de 1999 et 10 Gbits/pouce carré en 2002. Une limite connue avec les techniques actuelles est celle de l'effet superparamagnétique aux alentours de 30 Gbits/pouce carré. Il s'agit d'une perte de magnétisation de domaines trop petits. Toutefois, une nouvelle technique (2001) nommée medium antiferromagnétiquement-couplé (AFC) repousserait cette limite à 100 Gbits/pouce carré.

Une nouvelle technique propre à réaliser des mémoires centrales est annoncée sous le nom de MRAM (magnetic RAM) pour 2004-2005. Il s'agirait d'un dépôt magnétique pris en sandwich entre deux réseaux orthogonaux de conducteurs. Ses caractéristiques annoncées sont faible coût et grande capacité de la DRAM, grand débit de la SRAM, non volatilité de la flash, très faible consommation, bref, la merveille !

C6.2. La machine symbolique pendant les CINQ dernières années, éléments architecturaux

Ce thème a été abondamment traité dans les notes de cours, rappelons :

la montée en puissance des processeurs RISC;
 l'apparition des processeurs VLIW;
 la généralisation du pipeline;
 l'augmentation du nombre et de la taille des caches;
 la disponibilité des multiprocesseurs symétriques;
 la généralisation des bus multiples;
 l'arrivée des bus externes à grands débits;
 et, hors architectures, le foisonnement de la normalisation.

Pour ce qui est des tendances, on notera comme fait important l'évolution des processeurs RISC. Ils avaient été définis comme simples par rapport aux processeurs ordinaires des mêmes années. Aujourd'hui, les processeurs RISC sont faits pour utiliser le plus possible les capacités d'exécution simultanée. Ils fonctionnent en mode superscalaire, ont besoin de gérer l'exécution dans le désordre (out of order execution). On peut considérer que les MIPS R10000 et HP PA-8000 sont aussi complexes si ce n'est plus que les processeurs CISC ordinaires. Ils viennent à chauffer autant que les autres : 66 W à 1,3 GHz à 100 W à 2 GHz pour les Pentium 4, 109 W pour l'Alpha 21264 à 600 MHz sous 2,3 V.

C7. LES ÉVOLUTIONS PRATIQUEMENT ACQUISES OU LE COURT TERME, jusqu'à 2005.**C7.1. Les prochaines années de la machine physique.**

Il est banal de dire que les produits issus de l'électronique, de son voisin le magnétisme et de quelques autres techniques comme l'informatique vont continuer à évoluer rapidement dans le futur proche.

C7.1.1 En électronique.

Un tableau à la fois rétrospectif et prospectif a été publié dans [\[YU96\]](#). Il est intéressant parce qu'au lieu de faire une prévision seule, il compare les prévisions passées aux résultats acquis de 1989 à 2006. À ce titre il est un complément au [§ C6.1.1](#) ci-dessus.

Caractéristique	Prévision en 89 pour 96	Réalisation en 96	Prévision en 89 pour 2000	Prévision en 96 pour 2000	Prévision en 96 pour 2006
Nombre de transistors en millions	8	6	50	40	350
Taille en cm	2,32	1,78	3,05	2,79	3,56

Gravure en μm	0,35	0,35	0,2	0,2	0,1
MIPS	100	400	700	2400	20000
SPEC95 entier	2,5	10	17,5	60	500
Horloge en MHz	150	200	250	900	4000

Il est plus que vraisemblable que l'intégration des composants électroniques va se poursuivre pendant 10 ans au moins, certains, peut-être trop optimistes, disent 20 ans, même en l'absence de nouveautés radicales, par l'amélioration constante des techniques. Les prévisions, publiées ou non, sont à peu près établies.

On va voir en matière de circuits :

- la généralisation de la gravure à 0,25 μm , apparue dans le processeur PowerPC 750 à 300 MHz en 1998;
- son remplacement par la gravure à 0,18 μm apparue en 1999 [\[VER98\]](#);
- la gravure à 0,13 μm en 2002, annoncée à 0,10 μm en 2004.

La diminution de la taille de la gravure sera accompagnée des évolutions qu'elle nécessite ou qu'elle apporte :

- le remplacement de l'aluminium par le cuivre pour les conducteurs. Il apportera une dissipation plus faible et une mobilité électronique plus grande;
- l'abaissement de la tension d'alimentation à 1,5 V puis 1,1 V.

La fréquence de fonctionnement a atteint 1 GHz pendant l'année 2000, 2 GHz en 2001.

La dissipation thermique des circuits est liée à deux facteurs : la taille de gravure et la tension d'alimentation. Le tableau suivant contient son évolution à partir de la base arbitraire 100 pour 0,25 μm et 2,5 V. Le troisième facteur est la fréquence, la dissipation est **proportionnelle** à la fréquence.

Les consommations maximales de quelques Pentium 4 en 2001 illustrent concrètement les conséquences observées à gravure constante, 0,18 μm .

En fonction de la fréquence

66.68 W (1.3 GHz PGA423 @ 1.70 V)

71.05 W (1.4 GHz PGA423 @ 1.70 V)

75.25 W (1.5 GHz PGA423 @ 1.70 V)

83.98 W (1.7 GHz PGA603 @ 1.70 V)

97.24 W (2.0 GHz PGA603 @ 1.70 V)

En fonction de la tension :

097.24 W (2.0 GHz PGA603 @ 1.70 V)

100.45 W (2.0 GHz μ PGA478 @ 1.75 V)

Tension	gravure 0,35 μ m	gravure 0,25 μ m	gravure 0,18 μ m
2,5 V	100	41	n'est pas prévu
2 V	65	27	n'est pas prévu
1,8 V	n'est pas prévu	22	13
1,5 V	n'est pas prévu	14	11
1,1 V	n'est pas prévu	n'est pas prévu	8

Ce tableau a été publié dans l'Usine nouvelle n° 2655.

On peut donc espérer ici un gain maximal d'un facteur 10 à fréquence constante.

Ce tableau ne présente que des constats sans considérer deux phénomènes :

- le premier est relatif à la gravure, 0,13 μ m est atteint avec une lithographie en ultraviolet à 193 nm, il faudra 70 nm ultraviolet lointain ou par faisceaux d'électrons pour passer au dessous de 0,1 μ m et atteindre 8 GHz à technique CMOS constante;
- le second est que les déphasages inhérents à la transmission dans l'espace sont proportionnels à la fréquence.

Un attente raisonnable doit être fondée sur tous ces facteurs :

En conservant les coûts faibles de la gravure optique, des fréquences qui n'induisent pas des déphasages difficiles et coûteux à maîtriser et des dissipations acceptables pour des machines de bureau ou des machines portables, on pourra observer dans peu de temps : 220 millions de transistors en 0,18 μ m dans l'Itanium 2 d'Intel, fin 2002 ou début 2003 et 500 millions de transistors gravés en 0,13 μ m fin 2003 pour l'Itanium 3. Ces tailles sont celles de puces intégrant le cache L3 qui passera de 3 à 6 Mo.

Ces considérations optimistes ne concernent QUE la technique pure. La prise en compte de l'économie de la fabrication donne un éclairage instructif.

Résumé :

Approximativement en :	1966	1975	1990	1996	2003 (estimation)
La construction d'une usine de production de semiconducteurs coûte en milliards de francs	0,07	0,3	0,7	5	60
on y fabrique des puces dont l'aire est en cm ²	0,04	0,25	1,5	4	16
et dont la gravure est en μ m	15	6	1	0,25	0,1

Après trois années de stagnation relative de 1996 à 1998, le marché des semiconducteurs est en hausse depuis lors, 20% en

1999, 36% prévus en 2000 par Dataquest. La structure de la production était en 1999 en valeur : 40% pour les mémoires, 20% pour la téléphonie mobile et les usages vidéo numériques (télévision, lecteurs divers) et 40% pour les processeurs.

La concentration de la production paraît devoir se poursuivre. En 2010, on prévoit que les dix premiers fabricants réaliseront 90% des ventes contre 75% en 1999 (Jean-Philippe Dauvin, vice-président de STMicroelectronics octobre 2000).

IBM a annoncé en octobre 2000 la construction d'une usine pour un coût de 30 milliards de francs.

Intel a annoncé en novembre 2001 une nouvelle technique CMOS nommée TéraHertz pour une mise en service en 2005 avec pour objectifs la multiplication par 25 du nombre de transistors et la multiplication par 10 de la fréquence à consommation constante.

Il apparaît qu'Intel mise plus sur la fréquence et le nombre de transistors par puce alors qu'AMD mise sur l'augmentation du débit par améliorations architecturales.

C7.1.2. Pour les disques

Les évolutions attendues à court terme en densité d'enregistrement auront deux causes, une amélioration et un remplacement.

L'amélioration porte sur la densité d'enregistrement magnétique. On prévoit la généralisation de la technique de lecture dite [magnéto-résistive](#) qui portera la densité d'enregistrement à 10 Gbits/pouce carré en 2002.

Le remplacement pourrait être l'arrivée de certaines techniques magnéto-optiques. Seagate a présenté un disque Winchester à assistance optique ou OAW (optically assisted Winchester). Le principe mis en œuvre est l'effet Kerr. Il consiste à lire et écrire des orientations magnétiques par un faisceau de lumière cohérente polarisée. Les densités attendues sont d'au moins 10 Gbits/pouce carré.

Une autre réalisation semble-t-il proche, est celle de la société Constallation-3D. Elle propose, en premier lieu sous forme non réinscriptible, ensuite peut être réinscriptible des supports comparables aux CD-ROM. Leur nouveauté tient à plusieurs faits :

- .le support est transparent, constitué de tranches de polycarbonates;
- .chaque tranche est tartinée sur ses deux faces d'un matériau composite;
- .ce matériau est embossé;
- .les creux sont remplis d'un matériau fluorescent.

La lumière émise par fluorescence en réponse à un faisceau incident focalisé sur cette surface est ensuite détectée. Les caractéristiques attendues sont une capacité de 10 à 100 Go par empilement de tranches et des débits en lecture de 1 Go/s.

L'ensemble produit une division par 2 du [prix du volume](#) par an.

C7.2. Les prochaines années de la machine symbolique

La place disponible sur une puce et l'augmentation de la capacité de traitement apporteront des évolutions autant dans les architectures que dans les applications.

Dans les architectures, on verra

Une croissance du nombre et de la taille des caches qui contiendront des dispositifs de prévision inspirés de ce qui se fait déjà en matière de branchements.

Un élargissement des chemins d'instructions pour le VLIW, et moins certainement des *chemins de données* pour le multiprocesseur.

Une augmentation du nombre d'unités de traitement.

La profondeur des pipelines ne paraît plus devoir augmenter au delà des 20 étages atteints. Des dispositifs additionnels comme les [caches pour victimes](#) pourraient apporter des gains de l'ordre de 10%.

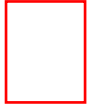
Les publications en 2000 ne contiennent pas d'annonce
de révolution comparable à celle du Risc.

L'espoir reste donc fondé sur la maîtrise du parallélisme. L'augmentation du nombre d'unités de traitement rendra les processeurs plus parallèles au niveau des données, prolongeant les techniques MMX et les réalisations superscalaires.

Dans un autre ordre d'idées, on a constaté que les progrès passés, qui étaient dans les mêmes ordres de grandeur que ceux que l'on prévoit, ont provoqué des bouleversements profonds dans les pratiques. Il y a encore 15 ans, on traitait du texte ou des nombres, l'affichage était alphanumérique et les imprimantes à aiguilles. On est passé de performances insuffisantes pour beaucoup des usages demandés, à un confort superflu qui a été utilisé pour faire naître de nouveaux besoins : images, sons, etc.

De la place sera disponible dans les tout prochaines années
pour de nouveaux types de traitements :
aide à la décision;
couplage d'applications;
gestion complète de la troisième dimension;
réalité virtuelle, etc.

Au delà du court terme, nous entrons dans des domaines où la dissipation d'énergie est la préoccupation majeure. On verra dans la suite deux directions de travail qui visent à la diminuer.




Conservatoire national des arts et métiers
Architectures des systèmes informatiques
HORS CHAPITRES
Conclusion, troisième partie
Année 2002-2003



Conservatoire national des arts et métiers
Architectures des systèmes informatiques
 HORS CHAPITRES
Conclusion, troisième partie
 Année 2002-2003



 Suite N°1...

Comment faire pour que, une fois sous tension, le circuit ne fonde pas ?
Question posée par l'auteur



C8. LES ÉVOLUTIONS MOINS CERTAINES OU LE MOYEN TERME

C8.1. Se garder des extrapolations

Le paragraphe précédent a fait la part belle aux extrapolations fondées sur les techniques acquises en laboratoire. Leur arrivée sur le marché est probable mais pas certaine. À plus long terme, le risque d'erreur en raison des incertitudes sur la technique et par ignorance de la demande.

Par expérience on sait tout de même quelques petites choses sur les innovations.

- Presque toutes les évolutions techniques et les diffusions d'innovations réussies ont une période de croissance exponentielle, c'est-à-dire un facteur multiplicatif constant par unité de temps.
- Toutes ont une période de décélération et parfois une diminution :
 - décélération du nombre de véhicules en service;
 - décélération du nombre de téléviseurs vendus, etc.;
 - diminution, jusqu'à 0 ou presque du nombre de machines à vapeur en service;
 - diminution du nombre de km de voies ferrées, etc.;
- En matière de transport le chemin de fer a fait passer la durée du voyage de Paris à Marseille de 12 jours à 12 heures en une fois, soit un facteur 24. Depuis le XIX^e siècle, cette durée a été réduite à 4 heures, soit un facteur 3. Le temps par avion de centre à centre est de toujours de l'ordre de 3 heures.

 Il n'est pas d'exemple qu'une technique ou une classe de produits
 ne connaisse pas une tendance
 à la saturation de la diffusion et des performances. 

C8.2 Le moyen terme de la machine physique.

There is a world-wide resurgence of interest in asynchronous logic design techniques. After two decades during which clocked logic has imposed its discipline across all corners of the world of digital logic, the older and more anarchic approach seems poised to make a come-back.

S.B. Furber, université de Manchester

On constate un retour mondial d'intérêt pour la logique asynchrone. Après deux décennies où la logique cadencée s'est imposée dans tous les recoins de la circuiterie numérique, cette pratique plus ancienne et plus anarchique paraît revenir.

C8.3 En matière électronique**C8.3.1 La gravure**

Les premières gravures à 0,08 μm ont été faites aux laboratoires Bell en juin 1996, reprises à l'université du Texas en février 1998. Le faisceau employé n'est plus de la lumière ultraviolette mais des électrons de plus de 100 keV. Le LETI à Grenoble a gravé un transistor de 0,02 μm en septembre 1999 avec des techniques encore optiques. Dans l'enthousiasme des réussites de laboratoires, on a pu lire des prédictions portant sur 16 Térabits par puce, 20 GHz et 100 Téraflopps.

La gravure industrielle à 0,13 μm est atteinte en 2002, 0,10 μm est attendu vers 2004.

La limite ultime de la gravure est estimée à 0,01 μm , soit une largeur de l'ordre de la centaine d'atomes. On est alors à la frontière du domaine quantique. Les électrons sont en nombre juste suffisant pour avoir un comportement statistique prédictible. Il y aura donc bien une saturation. Elle sera accompagnée d'une augmentation encore inconnue mais considérable du coût de fabrication.

Notons qu'en fréquence, Intel fait état de 10 GHz vers 2010.

On verra donc des progrès que l'on peut estimer par extrapolation, mais il faut aussi se tourner vers des procédés nouveaux ou que l'on croit nouveaux.

Dans les années 80, deux espoirs ont été déçus, au moins partiellement. L'auteur les signalait en ce temps là dans le cours d'architecture des systèmes comme des faits porteurs d'avenir.

Le premier espoir était fondé sur les composants en **arséniure de gallium** ou **AsGa**.

Quelques données relatives

au silicium (Si) et à l'arséniure de gallium (AsGa) à 300° K

Nom	Symbole	Unités	Si / AsGa
Structure cristalline			Diamant / Zinc Blende
Masse atomique	M	g. mole ⁻¹	28.09 / 144.63
Masse volumique	r	g.cm ⁻³	2.33 / 5.32
Nombre atomes par unité de volume	N	cm ⁻³	5.0 10 ²² / 4.4 10 ²²
Mobilité des électrons	μ_n	cm ² .V ⁻¹ .s ⁻¹ .	1.45 10 ³ / 8.5 10 ³
Mobilité des trous	μ_p	cm ² .V ⁻¹ .s ⁻¹ .	4.5 10 ² / 4.0 10 ²
Champ de claquage	E_{av}	kV.cm ⁻¹	3 10 ² / 3 10 ²
Coefficient de dilatation thermique	a_l	K ⁻¹	2.6 10 ⁻⁶ / 6.86 10 ⁻⁶
Chaleur spécifique	C	J.g ⁻¹ .K ⁻¹ .	0.7 / 0.35
Conductivité thermique	l	W.cm ⁻¹ .K ⁻¹ .	1.31 / 0.46
Température de fusion	T _f	° C.	1415 / 1240

Dans l'arséniure de gallium les électrons se déplacent plus rapidement que dans le silicium sous l'action d'un même champ électrique (mobilité des électrons). Ils commutent plus vite aux mêmes tensions, ou consomment moins à même rapidité par la réduction des tensions. Le gain en fréquence ou en puissance consommée est de trois à cinq.

Malheureusement, ils sont beaucoup plus chers. Le matériau ne s'obtient qu'en plaquettes de 7 à 10 cm au plus au lieu de 20 cm, ce qui divise par 4 le nombre de circuits intégrés de même taille fabriqués simultanément. Il est plus difficile d'y déposer des isolants et des métaux. En résumé, les méthodes de fabrication sont plus

ardues et les rendements sont plus faibles.

On n'utilise l'arséniure de gallium que s'il est indispensable, pour des fonctionnements à faible consommation au delà du gigahertz par exemple, évitant ainsi de passer d'un refroidissement par air à un refroidissement par fluide. L'utilisation des circuits intégrés en arséniure de gallium avait pris quelque essor dans les années 1980 pour des composants micro-ondes de transmission et des convertisseurs analogique-numérique; les super ordinateurs de Cray Computer Corporation, Cray 3 de 1989 à 4-5 Gigaflops puis Cray 4 avaient des composants en AsGa à 1 GHz. L'entreprise fondait elle-même ses circuits.

Remarquons que le gain de performance de 3 à 5 est rattrapé en une ou deux générations de circuits en silicium, ce qui oblige bien entendu l'industrie de l'arséniure de gallium à la même course à la performance que celle du silicium.

Leur résistance aux radiations peut être intéressante pour des usages dans l'espace.

Un processeur de type PowerPC a été fabriqué en 1995 à l'université du Michigan, gravure 0,5 µm, 400 000 transistors, 73 MHz. En 1996, les représentants de la même université annonçaient la possibilité de réaliser un processeur en AsGa complémentaire à 1 GHz. En matière de prix et de densité, les composants en AsGa ont de nombreuses années de retard sur les MOS et ne sont pas sur la voie de les rattraper.

En 1996, Texas Instruments a établi la faisabilité de mémoires statiques à effet tunnel (tunneling-based static random-access memories ou TSRAM) sur des supports de phosphore d'indium ou d'arséniure de gallium, avec les espoirs annoncés suivants :

- .consommation au repos 200 fois moindre que les SRAM;
- .des applications aux DSP en 2005 pour les radars;
- .des processeurs complets en 2010.

Ces produits ne sont pas dans le commerce.

Le second espoir était fondé sur l'**effet Josephson** découvert en 1962 par Brian Josephson à l'université de Cambridge. Il s'agit d'un "effet tunnel" qui peut se produire entre deux matériaux supraconducteurs séparés par un mince isolant. Dans un tel dispositif (nommé aussi "jonction Josephson"), les flux d'électrons, même en l'absence de toute différence externe de tension, peuvent passer d'un matériau à l'autre en franchissant l'obstacle isolant, au sein duquel il provoque une réaction magnétique d'une très grande sensibilité. Cette découverte est à l'origine de l'essor d'une technique nommée SQUID «superconducting quantum interference device» qui consiste à mettre en parallèle deux jonctions Josephson reliées en boucle par une liaison supraconductrice. Cet effet présent à très basses températures est apte à réaliser les fonctions des transistors. Les difficultés de mise en œuvre ont ruiné cet espoir. Le lecteur doit conserver en mémoire que la réalisation de la supraconductivité à hautes températures remettrait en course cet effet. La première conséquence serait la disparition de l'échauffement de ces composants.

À l'opposé on attend un grand retour, celui du fonctionnement asynchrone.

C8.3.2 La logique asynchrone.

La phrase en exergue de ce paragraphe mérite explication. Les deux qualificatifs synchrone et asynchrone ont des sens multiples en informatique. Pour les circuits, les significations sont claires, ce sont deux modes de fonctionnement d'un circuit combinatoire ou séquentiel.

- **Le mode synchrone** est celui d'un circuit commandé par une horloge qui distribue dans tout le circuit des battements réguliers. Les émissions de signaux et la commutations des circuits sont alignées sur ces battements, hors desquels il ne se passe rien. **Même s'il ne doit rien se passer, les circuits commutent... pour rien.**
- **Le mode asynchrone** est celui d'un circuit dans lequel il n'y a pas d'horloge. La commutation d'un composant est provoquée par l'arrivée d'un signal émis par un autre composant. Les composants qui ne travaillent pas sont au repos.

Analyse succincte du mode synchrone.

Les circuits numériques sont conçus comme l'agrégat de sous systèmes. Chacun est une machine à nombre fini d'états réglée par une horloge locale qui bat en permanence et dont le champ de commande définit l'étendue du

sous système. L'horloge locale est indépendante ou synchronisée sur une autre horloge. L'état du circuit est conservé dans des bascules, un nouvel état est calculé en fonction des entrées et de l'ancien état.

Ce fonctionnement a un aspect très favorable qui a été cause de sa généralisation.

- La conception et les tests des circuits sont considérablement plus simples. Ils vérifient chaque circuit et les délais d'acheminement des signaux. Ces techniques sont éprouvées. Les circuits sont vérifiables par des programmes de simulation en langages tels que VHDL. Mais tout ceci n'est vrai que si l'on se tient éloigné des fréquences et des distances qui introduisent des déphasages.

Ce fonctionnement a des conséquences défavorables ou génératrices de difficultés qui apparaissent aux fréquences très hautes et aux densités très grandes dont on approche.

- La consommation électrique est importante. Les nombreuses portes de pilotes d'horloges toujours en service commutent comme si tous les circuits commandés étaient en activité alors que seuls quelques uns sont actifs; parfois même aucun ne l'est. La technique CMOS a été retenue car ses portes ne dissipent de l'énergie que lors des commutations, mais dans une réalisation synchrone, toutes les horloges sont en permanence en état actif, qu'un signal les accompagne ou non. Les meilleurs microprocesseurs consomment et donc dissipent environ 30 watts.
- Les déphasages entre les signaux d'horloge deviennent difficiles à maîtriser car de plus en plus importants avec l'augmentation de la fréquence de fonctionnement.
- Des changements de fréquence existent quand un signal change de domaine d'horloge. C'est le cas des signaux qui viennent de l'extérieur du circuit comme les signaux d'interruption d'un ordinateur. C'est vrai aussi pour les signaux qui changent de sous système dans le même circuit. Un cache intégré de deuxième niveau fonctionne parfois à fréquence moitié de celle du reste de la puce.
- Il faut synchroniser les périphériques, mémoires et autres, à l'extérieur du circuit processeur. On le fait par les signaux des bus. En plus de la difficulté, cela contraint à aligner le fonctionnement de tous sur le plus lent dans les bus synchrones qui sont les plus rapides. C'est pourquoi on définit des retards systématiques sur le bus selon la mémoire installée.

La conception synchrone a été portée à son extrême dans les machines systoliques et aujourd'hui dans les machines RISC où non seulement chaque circuit est synchrone mais où le fonctionnement global veut l'être le plus possible. Cette conception approche ses limites, même si les informaticiens n'en ont pas la même conscience que les électroniciens.

La première critique forte du modèle synchrone date de 1985, par Chazelle et Monier [CHA85] qui montraient la nécessité de fonder une théorie générale de la physique du calcul. P. Vitanyi [VIT86] notait en 1986 que pour l'informaticien, un câblage a la propriété magique de transmettre instantanément d'un endroit à un autre alors que c'est la propriété de la seule machine symbolique. Dans un circuit synchrone, toutes les parties doivent être alimentées par l'horloge, idéalement de façon instantanée, du moins avec une tolérance de délais raisonnable.

Les difficultés de synchronisation entre circuits éloignés ont diminué avec l'unité centrale sur la même puce. Toutefois, même sur une puce unique, les déphasages des signaux d'horloge deviennent critiques quand la fréquence de fonctionnement augmente plus vite que la densité d'intégration en raison de la baisse de la tension d'alimentation. Dans les processeurs les plus rapides et dont la surface est de plus en plus grande, la partie de la puce consacrée aux pilotes d'horloge devient importante. On a envisagé sans grands résultats une transmission optique et non plus électrique pour diminuer les délais d'acheminement.

Les solutions ou les palliatifs.

Pour la consommation électrique. La diminution de la tension de fonctionnement réduit la dissipation d'un facteur 3 quand on passe de 5V à 3V et d'un facteur 6 quand on passe de 5V à 2V. Elle ne compense pas l'augmentation de la densité et l'élévation de la fréquence. Des processeurs qui consommeraient plusieurs centaines de watts ne pourraient pas être alimentés par batterie et demanderaient un refroidissement très coûteux pour des machines de table. Récemment (1999) un intégrateur a fait fonctionner des processeurs à 500 MHz dans des portables sans revoir le dispositif de ventilation. Les processeurs ont fondu.

Pour les signaux d'horloge. Quand la synchronisation globale est trop complexe et la commande par horloge consomme trop d'énergie, on pense à l'asynchronisme.

Les tout premiers ordinateurs ont été définis dans des styles divers. Deux machines de la première génération

étaient asynchrones, l'ORDVAC de l'université d'Illinois et l'IAS du groupe de Neumann à Princeton. Elles fonctionnaient dans les années 51-52. Plus tard le [MU5](#) de l'université de Manchester a fonctionné en 1974. L'idée avait été oubliée dira le pessimiste, avait été mis en attente dira l'optimiste.

La technique asynchrone **n'est pas simple**. Idéalement, l'arrivée des signaux de données met le composant en activité. Leur réception provoque un *éveil* du composant qui fonctionne alors à sa fréquence maximale. Le travail fini, le composant émet le résultat, revient à l'état de veille et ne consomme plus ou très peu d'énergie. Ainsi la consommation du circuit et les surchauffes diminuent considérablement. Les contreparties sont :

- la nécessité de «prévenir» le circuit de l'arrivée d'entrées;
- ma nécessité d'organiser les rendez-vous, substituts de la coïncidence des signaux d'horloge;
- l'obligation de prévenir des interblocages possibles.

Pour ce faire on utilise des protocoles de communication comme décrits ci-après. Ils prennent du temps et de l'espace. La complexité des circuits augmente.

Les travaux sur les processeurs asynchrones sont menés dans de nombreux endroits. Le programme européen Esprit finance le groupe de travail Asynchronous Circuit Design, Working group (ACiD-WG) depuis 1992. Il a treize participants :

Université d'Oxford, GB;
 Université South bank GB;
 Université cantabrique Espagne;
 Université technique du Danemark, Danemark;
 Université de technologie d'Eindhoven, Pays-Bas;
 Université de Groninge, Pays-Bas;
 Centre interuniversitaire de micro-Electronique, Belgique;
 Institut d'ingénierie des systèmes et des ordinateurs, Portugal;
 Université de Technologie de Loughborough, GB;
 Université de Manchester, GB;
 Laboratoires de recherche Philips, Pays-Bas;
 Université du Surrey, GB;
 Université polytechnique de Catalogne, Espagne.

Il finance également le projet OMI-MAP (Open Microprocessor systems Initiative - Microprocessor Architecture Project). Trois processeurs successifs nommés AMULET1 puis 2 et 3 [\[AMU98\]](#) ont été réalisés sur les plans du processeur ARM avec la technique micropipeline. Les travaux ont commencé en 1990 et se poursuivent. Ils ont prouvé que la consommation d'un circuit asynchrone équivalent à un circuit synchrone est nettement plus faible.

Des travaux sur les circuits asynchrones sont menés dans l'industrie, notamment par Sun Microsystems et Philips. On trouvera plusieurs références en bibliographie. Les outils de développement et de test des circuits asynchrones apparaissent. LARD, correspondant de VHDL est disponible gratuitement sur internet.

Souvent la conception asynchrone est vue comme une simple alternative au synchrone, alors qu'elle est multiforme. En voici quelques variantes.

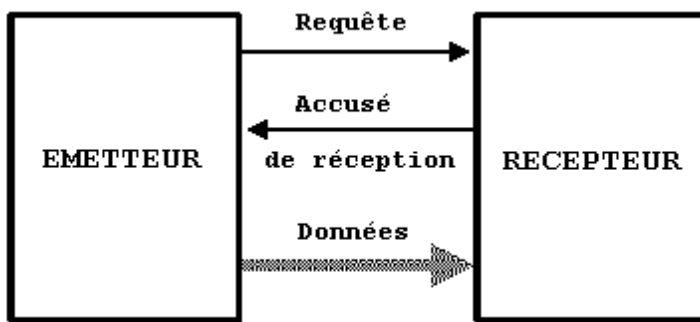
- **Sur les supports de codage du signal.** On a le choix entre deux types de codage. Le premier code chaque bit avec deux conducteurs, il est nommé codage de type 2n ou codage sur deux rails (dual rail encoding). Un porte la valeur, l'autre porte la disponibilité. Le second code chaque bit sur un conducteur et ajoute le signal de disponibilité de la botte de conducteurs sur un seul conducteur, il est nommé codage de type n+1 ou codage en botte (data bundling).
- **Sur le codage du signal ou le niveau face à la transition.** Les signaux codés en niveaux (level encoding) représentent le 1 logique par le niveau haut et le 0 par le niveau bas. Les autres (transition encoding ou event encoding) utilisent le changement pour porter l'information.
- **Sur le fonctionnement élémentaire.** On pratique deux types de conception. L'insensibilité aux délais (delay insensitive design) ne fait pas d'hypothèse sur les délais internes au système. Chaque porte ou connecteur peut prendre un temps quelconque pour commuter ou transmettre. L'indépendance à la vitesse (speed independent design) ne tient pas compte des temps de commutation des portes mais suppose que les transmissions sur les conducteurs sont instantanées.

On connaît aussi le codage de transition à deux conducteurs. Une transition sur l'un signale un 0, une transition sur l'autre signale un 1. Les niveaux n'ont pas de signification. Le circuit est alors totalement insensible aux délais de propagation.

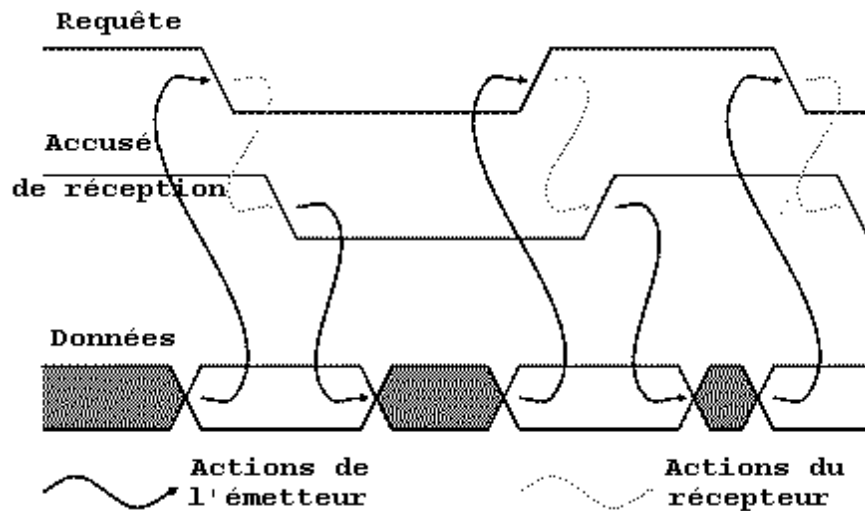
La forme la plus courante dans les maquettes et prototypes des dernières années est celle de deux conducteurs avec détection de niveau. Cette forme est toujours insensible aux délais mais il faut ménager une phase neutre entre les transitions. Les composants sont plus simples qu'en détection de transition. Ils sont aussi moins dissipatifs. Enfin, ils ne nécessitent pas l'enregistrement de l'état.

Un exemple de protocole de synchronisation est celui que Ivan Sutherland a présenté en 1987 dans son remerciement aux Turing Award. Il a présenté le «micropipeline». Les données sont en paquets, la transition est réalisée par un protocole à poignée de main qui commande le transfert des données. Dans cette conception, chaque bit utilise un conducteur et l'on ajoute non pas un mais deux conducteurs de disponibilité qui portent :

- la requête qui va de l'émetteur au récepteur, elle signale par une transition que les données sont disponibles;
- l'accusé qui va du receveur vers l'émetteur, il signale par une transition que les données ont été acceptées. L'émetteur retire les données quand il veut.



La séquence d'émission et réception est alors bien définie et le délai entre deux séquences est quelconque. Le protocole peut être représenté comme suit :



En résumé, les circuits asynchrones sont intrinsèquement plus complexes à concevoir et à tester en raison d'états instables et d'intervalles critiques.

Un produit de type asynchrone a été commercialisé par Sharp. Il s'agit d'un circuit spécialisé de traitement d'images et de décodage de type MPEG-2. La puce fait 3,8 milliards d'opérations par seconde, elle consomme 0,8 W, à comparer à la puce Alpha 21164 à 600 MHz de DEC qui fait 2,4 milliards d'opérations par seconde et consomme 109 W sous 2,3 V!

C8.3.3 Le stockage

La présentation des mémoires à holographie spectrale avec lesquelles on espère des densités de l'ordre de 8 Gbit/pouce carré est hors de notre propos. Le lecteur curieux consultera [\[HAI95\]](#).

C8.4. La machine symbolique

L'insuffisance d'algorithmes

En matière de logiciels d'application, certains auteurs croient déceler des limites de nature nouvelle. Elles tiendraient à notre méconnaissance des applications elles-mêmes. Autrement dit, on assisterait à une consommation sans renouvellement suffisant du stock de connaissances de modèles et d'algorithmes disponibles alors que les techniques et les coûts rendraient réalisables de nouveaux modèles et algorithmes. Cela semble être le cas dans les domaines suivants qui sont évoqués souvent comme champ d'utilisation des architectures parallèles :

- reconnaissance de la parole;
- phénomènes de catastrophes;
- modélisation du climat;
- modélisation des fluides turbulents (mécanique statistique des fluides turbulents et physique de l'atmosphère);
- gestion du code du génome des êtres vivants;
- circulation des océans;
- chromodynamique quantique en physique fondamentale, (étude de la propriété dite de «couleur» attachée aux quarks);
- etc.

D'ores et déjà, les capacités des circuits intégrés ne sont plus complètement exploitées alors que leur évolution ne ralentit pas.

Peut-être n'en sommes nous plus au comment le faire ? mais au «sais-je le faire ?»

La gestion de la complexité

Des méthodes et des outils de gestion de la complexité deviennent indispensables, fût ce au prix d'une baisse des performances.

Cette complexité devra être prise en charge. Les architectures, comme les modèles, les méthodes et les langages devront en prendre leur part. Les premières voies sont la modularité et l'abstraction qui ont déjà connu des succès considérables à la fois dans des logiciels généraux et dans la conception des circuits. La nouveauté appelée «middleware» comme le modèle de communication CORBA que l'on peut qualifier de bus logiciel, contribue à cette évolution [\[DUC98\]](#).

Une tendance profonde sera une progression simultanée :

- du matériel vers le logiciel actuel par l'architecture par la réutilisation des modules matériels;
- du logiciel vers le matériel par un accroissement de la capacité des compilateurs à utiliser finement le matériel disponible.

La part des architectures ne consistera pas en des inventions plus ou moins décousues mais en une reprise de signification. Une architecture est aussi une répartition de fonctions à réaliser en modules et un bon choix des mécanismes, plus ou moins abstraits, de communication entre eux. L'information doit être présente et disponible en un endroit, de préférence unique où on va la chercher et la mettre à jour après traitement. On gagnera à ce que les modules, si leur intercommunication est bonne, soient indépendants les uns des autres dans leur travail. La disponibilité de modèles uniques pour les différents transports pourrait apporter un gain en conception et en efficacité par une réutilisation alors possible, la [réutilisation](#) est apparue très tôt. Cet acquis sera conservé dans l'informatique quotidienne.

Conservatoire national des arts et métiers

Architectures des systèmes informatiques

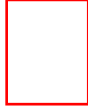
HORS CHAPITRES

Conclusion, troisième partie

Année 2002-2003



Conservatoire national des arts et métiers
Architectures des systèmes informatiques
 HORS CHAPITRES
Conclusion, troisième partie
 Année 2002-2003



Suite N°2...

One of the most remarkable attributes of mankind has been the ability to make progress in many technical fields despite the absence of a complete understanding of the underlying mechanisms. For example, we find ourselves in the midst of the electronic age and an electronic revolution, yet with little understanding of just what is an electron.

W.T. Grandy, Jr. Département de physique et astronomie, université du Wyoming

Un des caractères les plus remarquables de l'humanité est sa capacité à progresser dans de nombreux domaines techniques bien qu'elle ne comprenne pas complètement les mécanismes sous-jacents. Par exemple, nous sommes au milieu d'un âge de l'électronique et d'une révolution de celle-ci avec seulement une vague compréhension de ce qu'est exactement un électron.

C9. LES CHANGEMENTS SIMPLEMENT ESQUISÉS OU LE LONG TERME

C9.1 Sur les limites physiques du calcul.

Reprenons quelques idées éparses dans les paragraphes précédents.

Tout d'abord pourquoi cette note? Elle est utile dans la mesure où le fonctionnement des composants est loin du domaine sensible. Réduire d'un facteur deux une durée d'une seconde ou une surface d'un millimètre carré parle directement à l'entendement. Faire la même chose à une microseconde ou un nanomètre carré ne relève plus de l'expérience sensible.

La célérité des signaux. En tant que phénomène qui se propage, la célérité d'un signal ne peut pas dépasser la célérité de la lumière quelque soit le référentiel. (La lumière n'étant pas un objet mais une onde, nous employons célérité plutôt que vitesse. C'est l'explication du symbole c utilisé pour la désigner dans $E=mc^2$.)

La limite pratique est bien en deçà. La célérité du courant électrique est au mieux les deux tiers environ de la précédente dans un conducteur métallique. Les phénomènes de chargement et déchargement de capacités aux extrémités des conducteurs la diminuent encore. Une transmission à distance se fait au mieux à 100 000 km/s environ. Dans un ordinateur, c'est la compacité qui réduit les distances et apporte une diminution du délai sans augmenter la célérité.

La taille du semi-conducteur élémentaire. Cette taille limite est mal connue. La limite absolue est évidemment celle du plus petit cristal contenant au moins une impureté, mais ce cristal a un

fonctionnement quantique hautement imprévisible dans nos architectures. Il faut avoir une probabilité suffisante du fonctionnement recherché et pour cela que le nombre d'électrons soit tel que l'effet de moyenne opère. Une gravure à $0,02 \mu\text{m}$ correspond à une largeur de 200 atomes environ, soit 8×10^6 atomes pour un cube qui a cette arête. Le semi-conducteur peut aussi être fabriqué autrement : Des chercheurs de l'institut de physique théorique de l'université technique de Dresde (Saxe), en collaboration avec des chercheurs de l'institut Max-Planck de physique des systèmes complexes de Dresde ainsi que des chercheurs de l'université de Ratisbonne (Bavière) ont mis au point un transistor nanoscopique, cent fois plus petit que les commutateurs utilisés dans les microprocesseurs actuelles. Ces transistors sont constitués de nanotubes de carbones reliés entre eux par une molécule de fullerène (C_{60}). Par une intervention ciblée, la molécule de fullerène peut être pivotée de sorte qu'elle n'agisse plus comme passerelle bloquante entre les deux extrémités des nanotubes mais laisse au contraire passer le courant circulant à l'intérieur. La distance séparant les deux nanotubes ne dépasse pas le nanomètre.

Source : Dresdner Neueste Nachrichten, 25/02/01

Le nombre de connexions à partir d'une même porte. Cette limite est électrique. La puissance délivrée pour constituer un signal est limitée et le nombre de sorties d'une porte unique ne peut pas être augmenté au delà de quelques unités. C'est une limitation connue pour la taille des mémoires associatives et celles des piles câblées.

Tout ceci vaut pour une technique qui reste sur sa cinquième base, savoir l'intégration à grande échelle de composants Métal-Oxyde-Silicium ou de composants analogues en termes de semi-conducteurs. Les quatre premières bases étaient la mécanique, l'électromécanique, les tubes, les transistors isolés. Envisage-t-on des bases nouvelles ?

C9.2 Qu'en est-il d'autres voies ?

Une d'elles, que l'on peut qualifier de permanente parce qu'il en est question depuis longtemps sans jamais aboutir à une diffusion observée, est le calcul par des moyens optiques. Des transistors optiques ont été fabriqués, mais il n'apparaît pas que des innovations puissent intervenir avant au moins une dizaine d'années.

Plusieurs voies beaucoup plus nouvelles sont nommées :

- *nanotechnique* ou nanotechnologie ou encore *technique moléculaire*
- *calcul quantique*,
- *calcul réversible*.

Dans la technique courante, les effets sont fondés sur des commutations qui font intervenir des nombres considérables d'électrons et de molécules. Prenons l'exemple d'un circuit intégré pour lequel nous faisons les approximations suivantes : contenance 10 millions de transistors, masse un gramme, masse atomique moyenne des molécules le constituant 20, nombre d'Avogadro 6×10^{23} . On suppose que les conducteurs et les séparations occupent les 99 centièmes de la surface. Le nombre approximatif d'atomes de substrat par transistor est alors d'environ :

$6 \times 10^{23} (1/20) \times (1/1000000) / 100$ soit $6 \times 10^{23} / 2 \times 10^{10}$, ou encore 3×10^{13} .

Le nombre d'électrons en cause, proche du nombre d'atomes dopants, est de l'ordre de 10^{13} . L'objectif commun aux techniques moléculaire ou quantique est d'utiliser la molécule ou l'électron un à un ou au moins en très petit nombre. Le saut technique, même s'il restait partiel serait considérable. Toutefois, rien ne dit que ces techniques ne subiront pas le sort d'autres qui ont été attendues puis abandonnées.

C9.2.1 Nanotechnique

La ou les nanotechniques concernent ce qui peut être ultra miniaturisé comme les horloges, les robots et toutes sortes d'outils, ordinateurs inclus. Pour les ordinateurs, le modèle principal est biologique. On parle alors de calcul moléculaire qui tente de tenter d'imiter les fonctionnements des molécules comme les synapses des neurones ou d'autres [SCH99]. Nous ne sommes pas là dans le domaine quantique.

C9.2.2 Calcul quantique

Brièvement, le calcul quantique tente d'utiliser les propriétés essentielles du domaine quantique, savoir la superposition des états. Un ordinateur classique ou quantique manipule toujours des bits, systèmes physiques pouvant se mettre dans deux états distincts et représentant les valeurs logiques «oui» ou «non», 0 ou 1. Le bit informatique classique est égal soit à 0 soit à 1. Le bit quantique nommé qubit est simultanément dans les deux états. Par exemple, il peut valoir 0 avec une probabilité de 13 % et 1 avec une probabilité de 87 %. Un atome peut «tourner» simultanément dans les deux sens, un électron être présent sur deux niveaux d'énergie, un photon être polarisé dans deux directions, etc.

Deux bits ont quatre états distincts exclusifs (00), (01), (10) ou (11). Deux qubits ont ces quatre états simultanément, chacun ayant une probabilité, c'est la superposition des états. La modification d'un qubit affecte tous les états du système par le jeu des probabilités. Il y a un parallélisme massif qui pourrait faire atteindre des opérations hors de portée d'un ordinateur classique.

L'idée d'un ordinateur quantique vient de David Deutsch qui a montré en 1985 qu'un système à plusieurs qubits peut théoriquement servir de machine logique. L'essentiel est de savoir que ce système est théoriquement capable d'effectuer des opérations logiques tout à fait originales, avec une puissance de calcul considérable.

Le résultat d'un calcul fait sur ces innombrables états superposés serait obtenu par un contact avec le monde extérieur, une mesure par exemple, qui romprait la superposition. Les qubits prendraient la valeur 0 ou 1, suivant les probabilités définies par les lois quantiques. Actuellement une seule application, mais de taille, a été mise en évidence : la factorisation des entiers, c'est-à-dire la base du décryptage. Quant aux ordinateurs proprement dits, on connaît des formes plus que rudimentaires : une molécule à cinq atomes de fluor en suspension dans le chloroforme a fait un calcul quantique sur cinq qubits. des systèmes d'atomes de phosphore insérés sur une plaque de silicium, ou de molécules de fullerène coincées à l'intérieur de nanotubes sont essayés [HAR00].

C9.2.3 Calcul réversible

Le calcul réversible, quant à lui, concerne un effet qui n'est pas perceptible aujourd'hui, celui de l'irréversibilité entropique. Il mérite un développement.

Les lois des deux mécaniques classique et quantique sont réversibles. Le calcul tel qu'on le pratique est irréversible. La deuxième loi de la thermodynamique est une loi d'irréversibilité. Elle définit la croissance irréversible de l'entropie dans des conditions relativement larges. Pour le calcul, la limite entropique est un fait physique. Cette considération est distincte des techniques moléculaires qui comme leur nom l'indique ont pour objet de créer des circuits de nature différente.

Posons quelques jalons relatifs à ces notions peu souvent présentées et qui sont le seul pont entre physique et calcul. On dit parfois qu'il existe une physique du calcul.

La fait d'origine est une création de James Clerk Maxwell, celui qui a établi les lois de l'électromagnétisme. En 1871, il définit ce que l'on appelle depuis lors communément le démon de Maxwell. C'est une très petite créature qui voit et manipule les molécules une à une. Elle se tient sur un récipient empli de gaz, muni d'une barrière amovible en son milieu. Au début la barrière est ouverte et le gaz en équilibre. Les molécules sont toujours en mouvement puisque la température n'est pas le zéro absolu. Le démon ouvre et ferme la barrière laissant les molécules aller d'un coté à l'autre mais sans les laisser revenir. Alors, la pression augmente d'un côté et diminue de l'autre. Ce gaz comprimé pourra fournir un travail par détente vers l'autre moitié du récipient. Le démon paraît affranchi de la seconde loi de la thermodynamique. C'est un paradoxe. Il a été expliqué par L. Szilard [\[SZI29\]](#). Le système isolé n'est pas le récipient seul avec sa barrière et le gaz mais le récipient plus le démon. Alors l'entropie totale du récipient et de son gaz, et du démon et de son information reste constante. En 1929 L. Szilard [\[SZI29\]](#) imaginait le démon avec un récipient contenant une seule molécule. La présence de la molécule d'un côté ou de l'autre représente un bit d'information qui définit la quantité d'entropie minimale.

Une controverse assez longue a existé sur ce qui modifie alors l'entropie. Pour certains c'était le fait de mesurer l'information; pour d'autres, c'était le fait de l'effacer; pour d'autres encore c'était le fait de perdre la corrélation entre l'information détenue par le démon et l'état de la molécule. Parmi les étapes suivantes, Rolf Landauer [\[LAN61\]](#) a établi en 1961 **que seule la destruction d'information a un caractère dissipatif** irréversible.

Le lecteur qui a lu avec attention la distinction faite plus haut entre machine physique et machine symbolique a déjà l'attention en éveil sur l'imprécision des formulations précédentes. La molécule et l'information attachée à sa position sont deux choses différentes. Le fait de ne plus avoir de molécule ou de ne plus savoir où elle est a trait à la machine physique, au support de la donnée. L'information de position relève de la machine symbolique.

La conséquence des travaux des auteurs mentionnés plus haut est que le simple fait de faire des opérations irréversibles, par exemple ne garder que le résultat de l'addition de deux nombres en effaçant les opérands qui sont des bits, diminue l'information disponible et donc augmente l'entropie et par conséquent augmente la température du système.

On a vu une floraison d'études sur des modèles de calcul réversible. En 1963, Lecerf a établi que les machines de Turing réversibles sont universelles comme les autres. En 1973, Charles Bennett a montré que *tout* calcul peut être fait par pas réversibles. En 1975, Coppersmith et Grossman ont établi par la théorie des groupes, que les circuits réversibles dont la taille est d'un bit supérieure à la taille de leurs données d'entrée sont suffisants pour faire tous les calculs booléens. En 1977, Toffoli a montré que les automates cellulaires réversibles munis d'une dimension spatiale supplémentaire peuvent simuler les automates cellulaires non réversibles. D'autres résultats ont été établis dans les années 1980 et 1990.

L'auteur croit très immodestement avoir une vue sur le sujet, notamment par la distinction introduite entre la machine physique et la machine symbolique. Il n'apparaît pas clairement dans la littérature que l'information soit considérée en soi ou via le fait qu'elle n'existe que sous les espèces de sa réalisation en un phénomène. Si l'on accepte cette distinction, la destruction qui crée l'entropie n'est pas à mon sens celle du symbole qui disparaît ou est détruit, mais la destruction de sa représentation physique. Le lecteur peut alors dire : argutie ! puisque la condition nécessaire de présence du symbole est d'être écrit physiquement sous la forme d'un phénomène. L'auteur répond que ceci est bien exact mais que la distinction n'en demeure pas moins et que l'on peut dire que la destruction du symbole ne crée pas d'entropie et que la machine symbolique reste hors du domaine d'application de la physique. Si cet argument est valable, l'effacement ou pseudo effacement d'un fichier dans un système de Microsoft en remplaçant le premier caractère de son nom par une marque, crée l'entropie relative à l'effacement de 8 bits.

En contrepartie, l'effacement des secteurs du fichier augmentera beaucoup l'entropie. Et pourtant ceci n'est pas très clair. Si le nom du fichier n'est plus accessible, qu'est ce que cette information qui reste dans les secteurs et qui ne peut plus être connue?

De la même façon, si l'on considère les deux fonctionnements différents d'un support magnétique et d'un CD-ROM non réinscriptible. Dans le premier, il y aura destruction de l'information précédente par écriture en superposition, dans le second, l'information précédente reste intacte, même si on ne l'utilise plus, elle est remplacée par une autre information écrite ailleurs.

Note : dans toute la suite

k est la constante de Boltzmann soit $1,38 \times 10^{-23}$ joule par degré Kelvin,

T est la température absolue exprimée en °Kelvins, soit environ le °Celsius + 273.

ln2 est le logarithme népérien de 2.

Il y a réversibilité à deux conditions :

- Que l'on puisse reconstituer les entrées à partir des sorties, réversibilité logique.
- Que le dispositif matériel puisse fonctionner à l'envers, réversibilité physique.

La thermodynamique garantit alors qu'il n'y a plus de dissipation d'énergie due à la manipulation de l'information. Tout ceci ne vaut bien sur, que si d'autres consommations énergétiques ne viennent pas masquer ce phénomène. C'est le cas dans les circuits actuels et futurs qui seront toujours dissipatifs.

Quelle est la valeur de l'énergie dissipée par effacement, que ce soit le bit d'information, la donnée ou sa réalisation sur le support? Reprenons l'image du bit stocké en plaçant une molécule unique dans une boîte munie d'une séparation, formant une partie gauche et une partie droite, système à deux états. L'effacement consiste à placer définitivement la molécule dans une partie de la boîte, sans considération de l'endroit où elle se trouvait auparavant. Ceci peut être fait en supprimant la séparation et en comprimant le gaz formé de cette seule molécule de telle sorte qu'elle ne puisse être que dans le côté choisi. Cette action réduit l'entropie du gaz de $\Delta S = k \ln 2$ et le flux de chaleur associé s'échappe de la boîte. Si le processus est isotherme à la température **T**, un travail $W = kT \ln 2$ est accompli dans la boîte. L'effacement ou destruction d'information est payé.

Note intermédiaire : L'attention du lecteur est attirée sur les distinctions suivantes. La dissipation ainsi calculée est celle qui est inhérente au calcul et non celle qui est due à la machine physique et à ses processus propres. De la même façon qu'on peut toujours utiliser un marteau pilon comme chasse mouche, la quantité dissipée est pour l'essentiel aujourd'hui celle du processus physique utilisé. La limite théorique due à l'irréversibilité du calcul ne deviendra perceptible que dans un futur au moins éloigné.

Alors, quel est cet éloignement ? Les dissipations actuelles sont de plusieurs ordres de grandeur au dessus de la limite théorique des circuits destructeurs. Mais, si l'on prolonge la tendance actuelle à la diminution de la taille du transistor et donc la diminution de sa dissipation, la visibilité de la limite kT pourrait être atteinte vers 2015. Il s'agit donc de prendre date pour un futur que l'on sent proche.

Les tenants du calcul quantique estiment que l'une des trois possibilités suivantes se produira :

- ou la décroissance actuelle de la diminution de dissipation ralentira et s'arrêtera dans les vingt prochaines années,
- ou les ordinateurs fonctionneront à des températures basses,

- ou on mettra en service des ordinateurs nouveaux réversibles pour échapper à la limite kT .

Le lecteur remarquera que les arguments attachés aux circuits asynchrones, savoir que l'on peut dès maintenant diminuer la dissipation en ne faisant fonctionner que les circuits utiles, n'est pas vraiment prise en compte dans ce raisonnement.

La première possibilité n'est pas vraiment enthousiasmante. Les acrobaties techniques de Control Data pour refroidir et réguler le 6600 par circulation forcée de fréon dans les années 60, et par S. Cray dans les Cray 3 et 4 montrent à l'évidence que la dissipation par porte restera, en l'absence d'alternative, une limitation forte de la puissance des ordinateurs. Des bons esprits estiment que la consommation des systèmes informatiques absorbe 5 % des ressources électriques aux États-Unis et que les 10 % pourraient être atteints au début des années 2000.

La deuxième éventualité ne diminue pas la dissipation d'énergie. Supposons un instant que l'on opère à $3^{\circ}K$, la dissipation intrinsèque sera diminuée d'un facteur 100 mais il faut prendre en compte le réfrigérateur dont la performance thermodynamique sera de l'ordre de $3/(300-3)$ c'est à dire 0,01%. Le gain énergétique est largement compensé par la dépense nouvelle.

La troisième paraît la plus séduisante. Pour cela il faut envisager de fabriquer des circuits réversibles.

Les portes logiques utilisées ordinairement sont irréversibles :

A	B	NON A	NON B	ET	OU	OU EXCL	NON ET
0	0	1	1	0	0	0	1
0	1	1	0	0	1	1	1
1	0	0	1	0	1	1	0
1	1	0	0	1	1	0	1

Seuls les opérateurs NON A et NON B autorisent la réversibilité.

Les autres ne permettent pas de retrouver les deux bits d'entrée à partir du bit de sortie. Selon la loi de Landauer le travail minimal nécessité par les quatre dernières est $W = kT \ln 2$.

Nous pouvons construire des opérateurs réversibles en utilisant les portes de Toffoli à trois entrées et trois sorties, par exemple le ET est construit avec

(a,b,c=0) à (a,b,c=ET(A,B))

Le troisième bit est changé si les deux premiers valent 1 et sinon il reste à 0.

Le lecteur va dire qu'à l'issue du calcul complet, il faudra bien détruire les données rendues inutiles. Le théoricien répond qu'à l'issue du calcul, toutes les opérations seront refaites à l'envers, revenant à la situation initiale de façon parfaitement réversible.

Ces considérations ne sont plus tout à fait une vue de l'esprit. Un périodique sérieux existe nommé Nanotechnology et des congrès réguliers sont tenus, le cinquième en novembre 1997.

The Fifth Foresight Conference on Molecular Nanotechnology.

On en trouvera les résumés des conférences dans :

<http://www.ai.mit.edu/~mpf/Nano97/abstract.html>

Les travaux actuels portent donc à la fois sur la définition de calcul réversible, celui qui évite la destruction d'information génératrice d'entropie et de chaleur, et sur les techniques de fabrication.

On y trouve des suggestions de circuits logiques extrêmement petits, rapides, et dont la consommation énergétique serait infime. Ils seraient la base de machines asymptotiquement réversibles, exécutant des opérations logiquement réversibles.

Hélas, il n'est pas prouvé que les nouveaux composants en cours de définition se révéleront parfaitement fabriquables et utilisables.

Les considérations ci-dessus sont loin d'être du domaine concret quotidien. Il y a quand même espoir que les informaticiens aujourd'hui trentenaires, utilisent de telles machines à la fin de leur vie professionnelle.

The program to re-investigate the fundamental principles of physics from the standpoint of information theory is still in its infancy.

A.M. Steane, Reports on Progress in Physics, vol 61, 1998.

Le réexamen des principes fondamentaux de la physique du point de vue de la théorie de l'information est encore dans l'enfance.

Une réalisation illustre à la fois les difficultés et les espoirs que l'on peut avoir :

En 1989, IBM a réalisé une communication sécurisée fondée sur des propriétés de physique quantique. L'objectif était de transmettre une clef de chiffrement sans qu'elle puisse être interceptée. Pour cela, l'expéditeur a envoyé des photons au destinataire. Selon les règles de la physique quantique, l'état de ces photons ne peut pas être observé sans qu'il soit modifié. Une interception de la communication pourra ainsi être détectée par le destinataire qui peut avertir l'expéditeur d'arrêter la transmission. La portée de la transmission était de quelques dizaines de centimètres mais on sait transmettre aujourd'hui des clés de chiffrement à plusieurs dizaines de kilomètres dans des canaux quantiques.

Il nous faut mentionner enfin les machines (très futuristes) qui seraient basées sur le fonctionnement de chaînes d'ADN. Les organismes impliqués sont entre autres l'université de Turku en Finlande, les laboratoires de NEC, le Caltech et l'université de Princeton. Le lancement de cette idée date de 1994. Leonard Adleman, mathématicien américain a trouvé un moyen de résoudre le problème bien connu en recherche opérationnelle du voyageur de commerce pour 7 villes. Pour faire bref, Adleman a fabriqué les fragments d'ADN qu'il a jugé nécessaires, il leur a appliqué les enzymes d'accélération de la réaction et a obtenu une semaine plus tard la chaîne d'ADN codant pour la solution. On en a conclu abusivement qu'il y avait là un ordinateur à ADN. Cependant, cette technique peut éventuellement apporter une solution à des problèmes dont la combinatoire est considérable et hors de portée du calcul séquentiel. L'ADN ne fait pas des milliards d'opérations par seconde mais peut avoir une capacité à traiter des problèmes dont on ne connaît pas l'étendue. Les optimistes écrivent que l'on pourrait faire des calculs complexes qui demanderaient des milliers d'années à des machines parallèles et que les capacités de stockage seraient considérables. Une seule molécule d'ADN pourrait fournir plus de mémoire que tous les ordinateurs actuels n'en contiennent.

On trouvera des références sur ces sujets en bibliographie et le lecteur peut toujours consulter les bibliothèques et la toile pour son propre compte.

Enfin, et de façon plus fondamentale, la question est toujours posée de la limite du schéma de Turing et Church qui a repris les idées de Russell :

Le raisonnement déductif, l'algorithmique,
épuise-t-il le mode de raisonnement humain ?

Conservatoire national des arts et métiers
Architectures des systèmes informatiques
HORS CHAPITRES
Conclusion, troisième partie
Année 2002-2003

