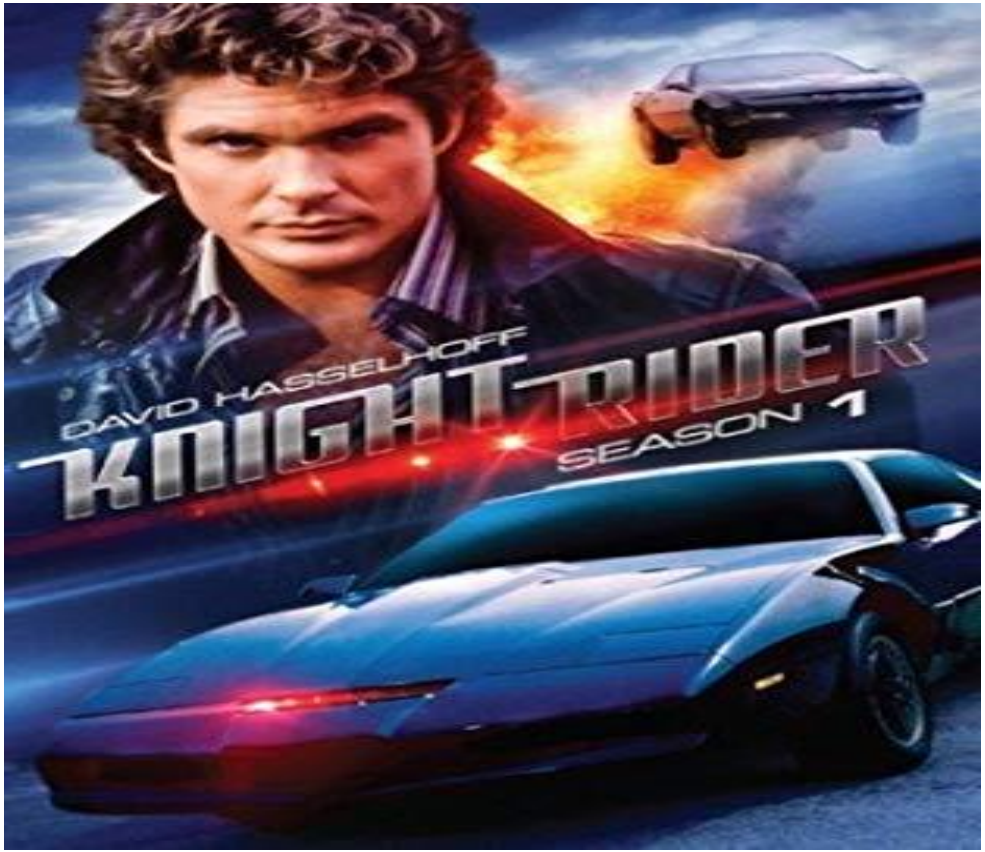
Accelerated Computing Solutions for AI and HPC Workloads

Sarosh Irani

2019

Knight Rider (1982)



The Future has Arrived

- Personal Assistants
  - Alexa, Siri, Google Assistant, Cortana, Bixby, Watson

- Self Driving Cars

Two Trends driving Computer Architecture

- Slowdown in Moore's Law
- Growth of Cognitive Computing
  - Machine Learning and Neural Network based computing

No more Moore?

Moore's Law:

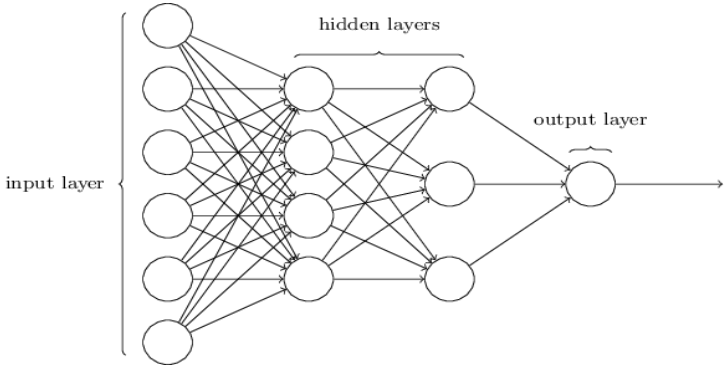
- Trend observed by Gordon Moore in 1965
- "Number of transistors on a silicon die doubles every 2 years" – predicted to continue for a decade
- Sometimes quoted as performance doubles every 18 months (accounting for more and faster transistors)
- Basis of the Tick-Tock model that Intel has been executing on for last 2 decades, though breaking down now
- Has run longer than most experts imagined, but now running up against the laws of Physics
  - From 2,500 transistors to 25 Billion transistors (7 orders of magnitude)

Denard Scaling:

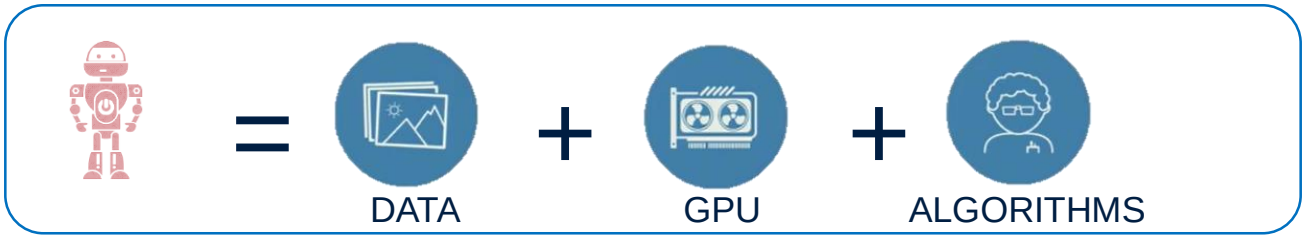
- When scaling down to a smaller node, voltage and current also scale down – chip supply voltages scaled down to under 1V
- Ended around 2005
  - Frequency race moved us quickly from 100 Mhz to 3 GHz, but we have been approximately flat since then

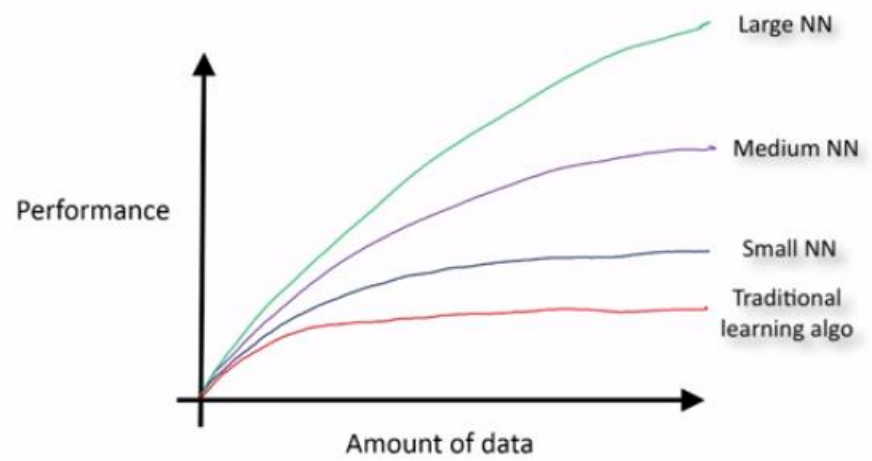
Neural Network based Computing

- Alternative approach to current 'Algorithmic Computing'
  - Traditional Compute adequate for solving many problems, enabled putting a man on the moon
  - Neural Network Computing is Stochastic not Deterministic
- Sometimes also referred to as 'Cognitive Computing'
- Very promising results achieved in areas that had proved hard for traditional compute
  - Image recognition, speech recognition, language translation etc



Deep Learning on Neural Networks – why now?



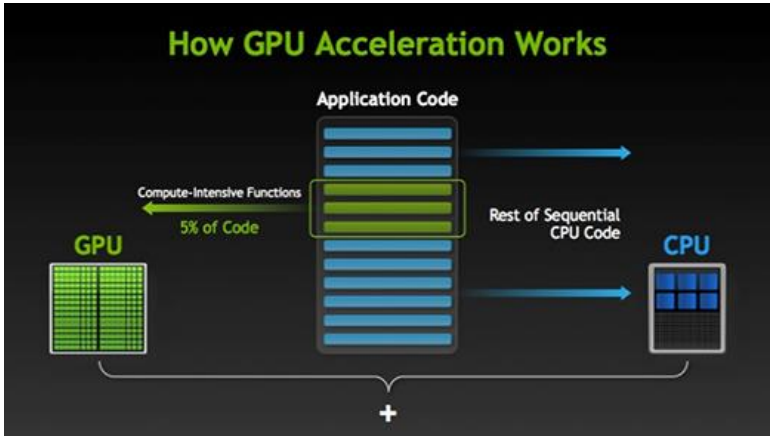


AI the Killer App

- When technology keeps increasing exponentially, we need to find a reason to leverage it
  - No one cares for a 1,000 mph car, cause you don't have roads to drive it on
- Killer app – new highly desirable feature/application that is hard to run on current computational systems
- Some examples from the past few decades
  - Microsoft Office
  - Internet (driving connectivity speeds higher than 28K, 56K)
  - High Quality Video
- AI is the killer app for hardware today

Increasing role of GPU in Performance Computing

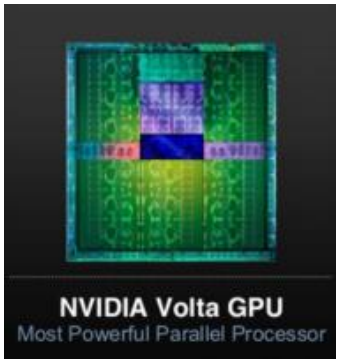
- GPUs, DSPs, ASICs, FPGAs have all been around, but presently we are seeing a strong trend to having compute intensive workloads migrate from CPU to GPU
- **Key drivers**
  - CPU frequency no longer scaling in the last 15 years (due to end of Denard Scaling)
  - More focus and emphasis on performance gains thru Parallel Processing instead of faster clock speeds – rise of multithreading, multi-core processing
  - GPU is a massively parallel computing device (5,000+ CUDA cores on Nvidia V100)



  - *Many Top 500 Supercomputers today follow this paradigm and are 'GPU Accelerated'*

X11 GPU Server Family

- **This is our Fourth Generation of specialized servers for Parallel Computing**
  - X11 supports Skylake/Cascade Lake and Volta, Turing GPU families
- **Largest GPU server portfolio in the industry**
  - Strong year over year growth; continued investment in expanding our product line
- **Sell into numerous Parallel Compute, HPC Verticals**
  - Oil & Gas, CAD/CAE, Computational Finance, Research & National Labs, Hyperscale Cloud
- **With Deep Learning impacting numerous industries, we see a much larger TAM**



X11 GPU Server Portfolio

Ratio: GPU:CPU

GPU OPTIMIZED

Tower/4U

7049GP-TRT
4:2 (4U)

Rack – 1U/2U

1029GQ-TRT
4:2 (1U)

1029GP-TR
3:2 (1U)

1019GP-TT
5019GP-TT
2:1 (1U)

1029GQ-TVRT
4:2 (1U)

2029GP-TR
6:2 (2U)

Rack – 4U/10U

6049GP-TRT
20:2 (4U)

4029GP-TRT
8:2 (4U)
Dual Root

4029GP-TRT2
10:2 (4U)
Single Root

9029GP-TNVRT
16:2 (10U)

4029GP-TVRT
8:2 (4U)
NVLink

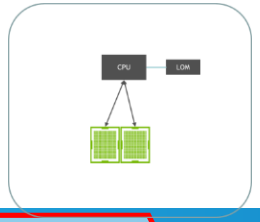
X11 Parallel Computing Servers

Best–in-class technology designed for highly parallel applications to deliver ultimate performance, flexibility, and scalability

2 1019GP-TT/5019GP-TT

Cost Effective

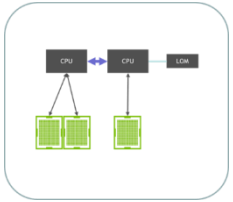
- UP SKYLAKE CPU
- 6x 2.5” HS HDD bays (1019GP-TT)
- 3x 3.5” HS HDD bays (5019GP-TT)
- 2 Double-Width GPUs
- 1 x16 PCIe 3.0 slot
- 1x 1400W Platinum PWS



3 1029GP-TR

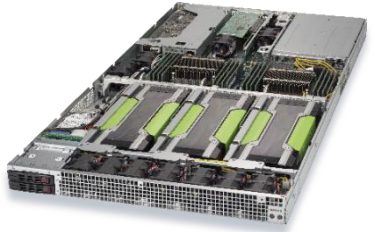
Flexibility

- DP SKYLAKE CPU; 3UPI
- 4x 2.5” HS HDD bays
- 3 Double-Width GPUs
- 1 x16 PCIe 3.0 slot, SIOM
- 2x 1600W Platinum PWS

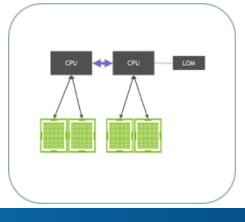


4 1029GQ-T(N)RT

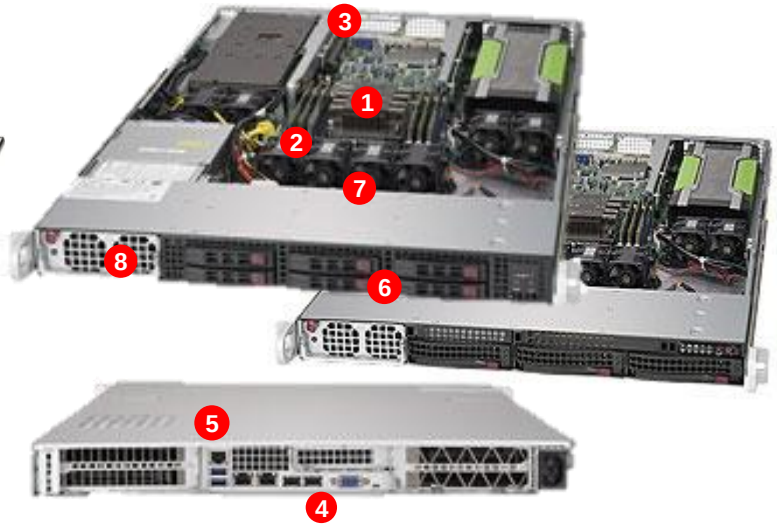
Performance

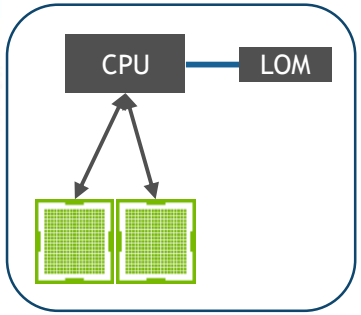


- DP SKYLAKE CPU; 3 UPI
- 4x 2.5” HS HDD bays; NVMe
- 4 Double-Width GPUs
- 2 x16 PCIe 3.0 Slots
- 2x 2000W Titanium PWS



SYS-1019GP-TT/5019GP-TT





Key Features:

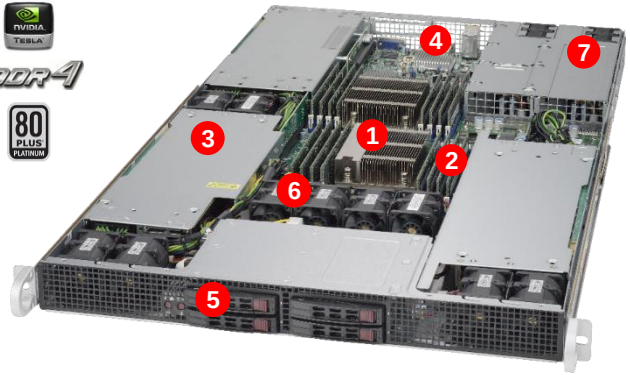
- Entry level offering
- Single CPU, directly connected to 2 GPUs
- Support for 2.5” & 3.5” drives

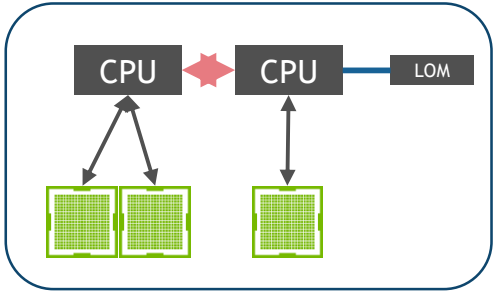
Key Applications:

- Oil & Gas/Seismic
- Scientific/Data Mining

1. **Processor Support**
   Single Xeon Scalable Processor (Skylake)
2. **Memory Capacity**
   6 DIMM ECC DDR4 2666 MHz
3. **Expansion Slots**
   2 PCI-e x16 Gen 3.0 for double-wide GPU cards
   1 x16 Gen 3.0 LP card
4. **I/O ports**
   1x VGA, 2x GbaseT LAN, 2x USB 3.0, and 1x IPMI dedicated LAN port; 2x M.2 NVMe
5. **System Management**
   On board BMC (Baseboard Management Controllers) supports IPMI2.0, media/KVM over LAN. (Dedicated LAN port for management)
6. **Drive Bays**
   1019GP: 6 Hot-Swap 2.5” Drive Bays
   5019GP: 3 Hot-Swap 3.5” Drive Bays
7. **System Cooling**
   8 counter rotating fans w/ optimal fan speed control
8. **Power Supply**
   1400W Platinum Level efficiency power supply

SYS-1029GP-TR





1. **Processor Support**
   Dual Xeon Scalable Processor; 3 UPI
2. **Memory Capacity**
   16 DIMMs ECC DDR4 2666 MHz
3. **Expansion Slots**
   3 PCI-e x16 Gen 3.0 for double-wide GPU cards
   1/1 x16/x8 in LP slot
4. **I/O ports**
   1x VGA, SIOM support, 2x USB 3.0, and 1x IPMI dedicated LAN port
5. **Drive Bays**
   4 hot-swap 2.5” drive bays
6. **System Cooling**
   10 counter rotating fans with optimal fan speed control
7. **Power Supply**
   1600W Platinum-Level efficiency redundant power supply

Key Features:

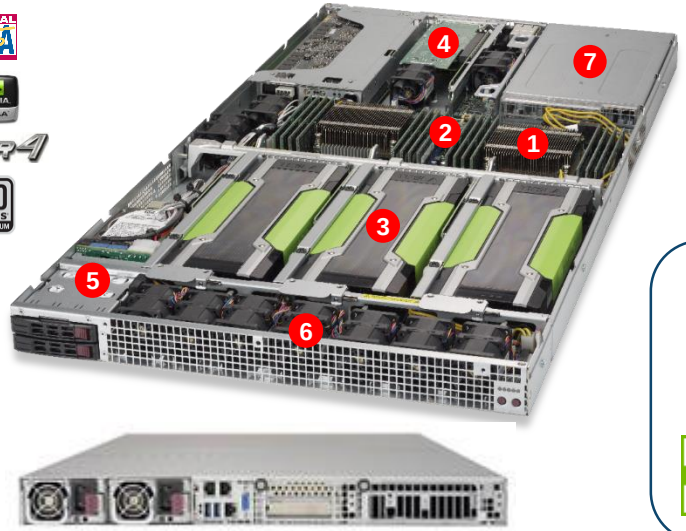
- Dual CPU with 3 GPUs
- Support for 16 DIMMs
- SIOM supported
- 1600W Platinum Power Supply

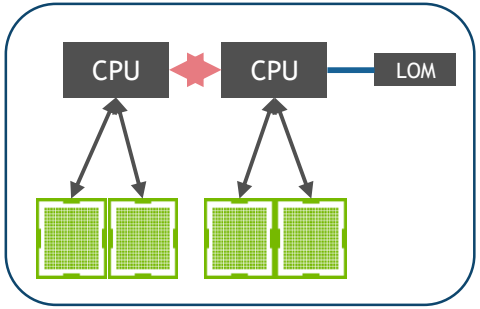
Key Applications:

- VDI technology
- HPC
- Machine Learning
- Computational Finance

SIOM

SYS-1029GQ-T(N)RT





Key Features:

- 4 Tesla V100 GPUs in a 1U
- Support for active and passive cooling
- 2000W Titanium power supply

Key Applications:

- Oil & Gas
- Research & Scientific
- VDI technology
- Computational Finance

1. **Processor Support**
   Dual Xeon Scalable Processor; 3 UPI
2. **Memory Capacity**
   12 DIMMs ECC DDR4 2666MHz
3. **Expansion Slots**
   4 PCIe3 x16 for double-wide GPU cards
   -TRT: Two x16 LP card
   -TNRT: x16/x8 LP card
4. **I/O ports**
   1x VGA, 2x 10GbaseT LAN, 2x USB 3.0, and 1x IPMI dedicated LAN port
5. **Drive Bays**
   -TRT: 2x HS 2.5” SATA drives bays; 4x total 2.5” HDD bays
   -TNRT: 2x HS 2.5” NVMe drives bays; 4x total 2.5” HDD bays
6. **System Cooling**
   9x counter rotating fans with optimal fan speed
7. **Power Supply**
   2000W Titanium redundant power supply

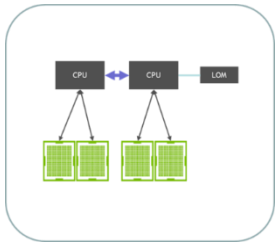
X11 Parallel Computing

Best–in-class technology designed for highly parallel applications to deliver performance, flexibility, and scalability

4 7049GP-TRT

Workstation

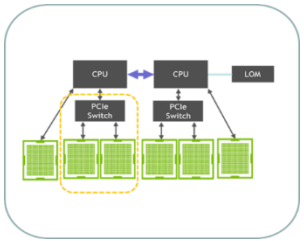
- DP Skylake CPUs
- 8x 3.5” HS HDD bays
- 4 Double-Wide GPUs
- 6 x16 PCIe3 slots
- 2x 2000W Titanium PWS



6 2029GP-TR

Mainstream

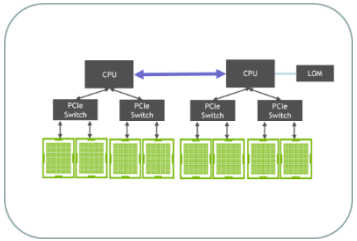
- Dual Skylake CPUs
- 8x 2.5” HS HDD bays
- 6 Double-Wide GPUs
- 1 x16 PCIe3 slots, SIOM
- 2x 2000W Platinum PWS



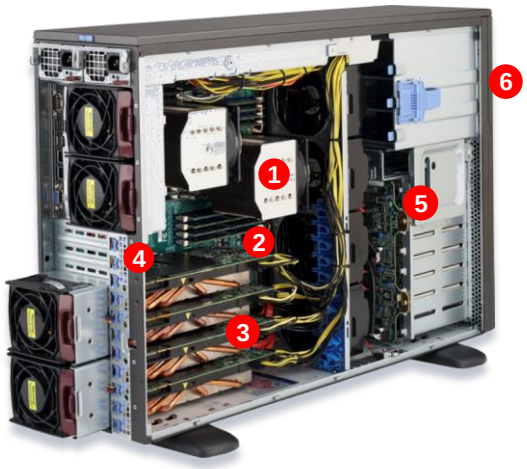
8 4029GP-TRT

Parallel Optimized

- Dual Skylake CPUs
- 24x 2.5” HS HDD bays
- 8 Double-Wide GPUs
- 2 x16 PCIe3 slot;
- 4x 2000W Platinum PWS



SYS-7049GP-TRT



4U Rackmount Option with
MCP-290-00059-0B (rail) and
MCP-210-74703-0B (bezel)

Key Features:

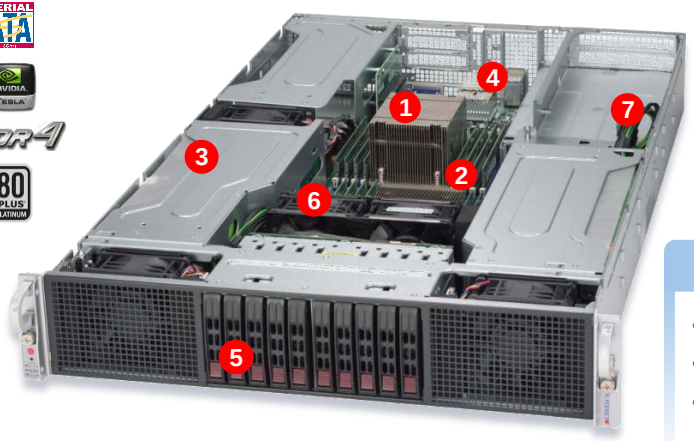
6 PCIe3 x 16 supporting 4 DW GPUs
Optional four NVMe drives support
NVMe/SAS3 backplane
Dual 10GbT onboard

Key Applications:

Research & Scientific
Simulation and Creation Design
Computer Aided Engineering
Machine Learning

1. **Processor Support**
   Dual Xeon Scalable Processor; 3 UPI
2. **Memory Capacity**
   16 DIMMs ECC DDR4 2666MHz
3. **Expansion Slots**
   4 PCI-e 3.0 x16 for double-width GPU cards,
   2 PCI-e 3.0 x16 for PCIe add-on card
   1 PCI-e 3.0 x4 (in x8)
4. **I/O ports**
   1x VGA, 1x COM, 2x 10GbE LAN, 4x USB 3.0, 2x USB 3.0, and 1x IPMI dedicated LAN port, Audio 7.1
5. **System Cooling**
   4 heavy duty fans, 4 exhaust fans, and 2 active heat sink w/ Optimal Fan Speed Control
6. **Power Supply**
   2200W Titanium Level efficiency redundant power supplies with DC240V support

SYS-2029GP-TR

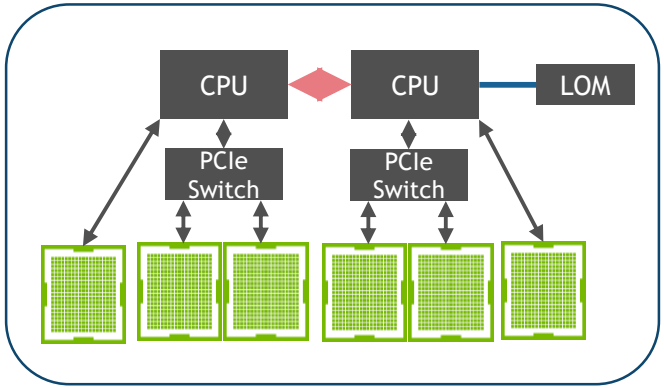


Key Applications:

- Computational Finance
- Oil and gas
- Weather and Climate Analysis

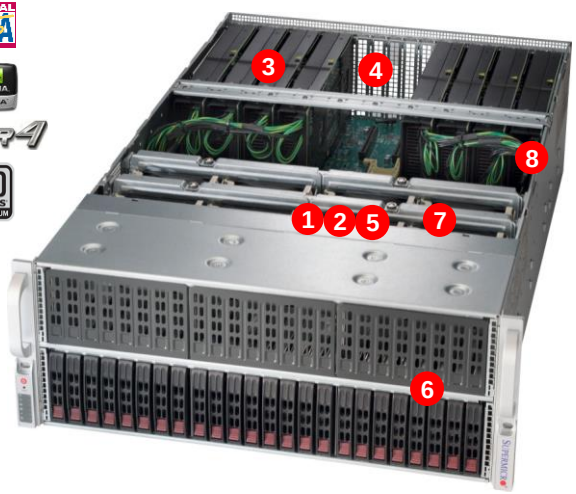
Key Features:

- 6 GPUs in a 2U
- 10 hot swap 2.5” drive bays
- SIOM support
- 2000W Platinum power supply



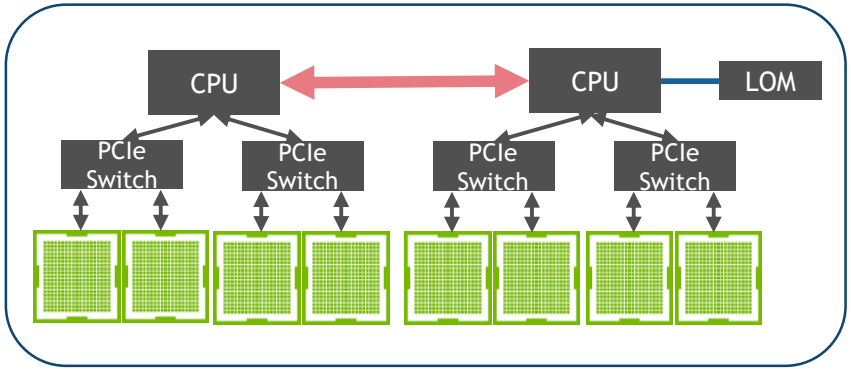
1. **Processor Support**
   Dual Xeon Scalable Processor; 3 UPI
2. **Memory Capacity**
   16 DIMMs DDR4 2666 MHz
3. **Expansion Slots**
   6 PCI-e x16 Gen 3.0 for double-wide GPU cards;
   1/1 x16/x8 in LP slot
4. **I/O ports**
   1x VGA, SIOM support, 2x USB 3.0, and 1x IPMI dedicated LAN port
5. **Drive Bays**
   10 hot-swap 2.5” drive bays
6. **System Cooling**
   5 counter rotating fans with optimal fan speed control;1 air shroud
7. **Power Supply**
   2000W Platinum Level efficiency redundant power supply

SYS-4029GP-TRT



Key Features:

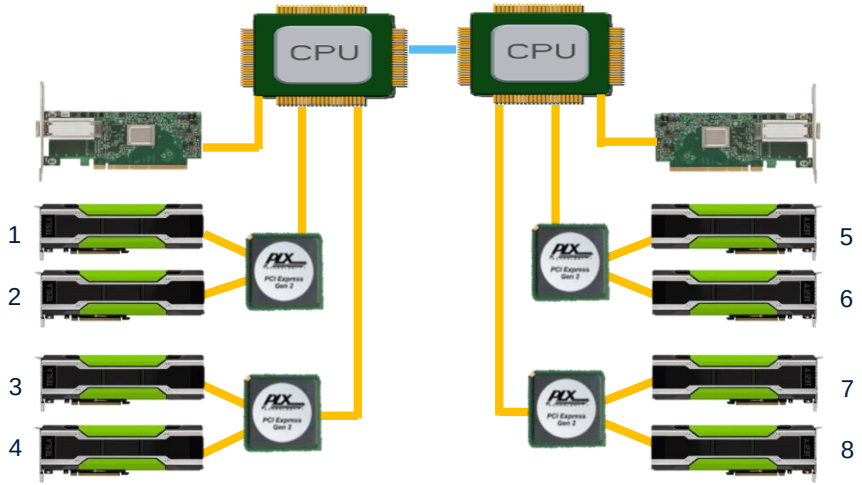
- Supports 8 double wide GPUs
- Up to 24 hot swappable 2.5’ drives
- 4 x 2000W Platinum Power Supplies



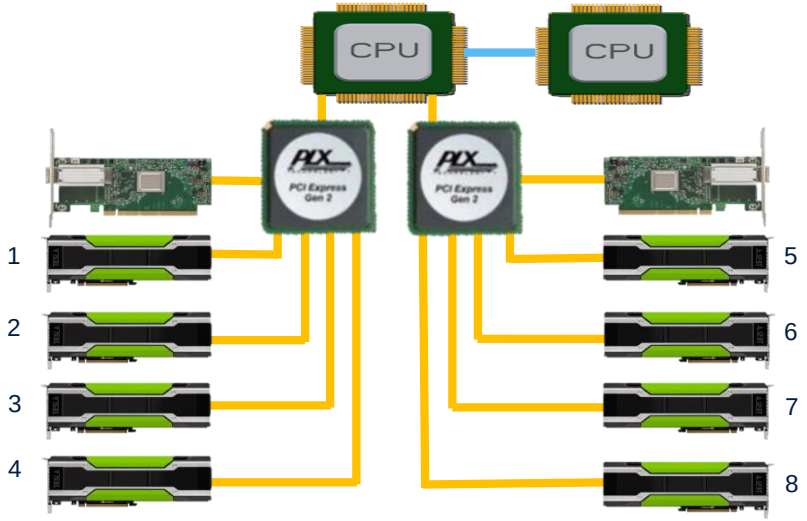
1. **Processor Support**
   Dual Xeon Scalable Processor; 3 UPI
2. **Memory Capacity**
   24 DIMMs ECC DDR4 2666 MHz
3. **Expansion Slots**
   8 PCIe 3.0 x16 for double-wide GPU cards
   2 PCIe 3.0 x8 (2 in x16 slots)
   1 PCIe 3.0 x4 (in x16)
4. **I/O ports**
   1x VGA, 2x 10GbaseT LAN, 4x USB 3.0, and 1x IPMI dedicated LAN port, 1x M.2 NVMe
5. **System management**
   On board BMC (Baseboard Management Controllers) supports IPMI2.0, media/KVM over LAN with dedicated LAN for system management
6. **Drive Bays**
   24 hot-swap 2.5” drives bay
7. **System Cooling**
   8 heavy duty fans optimize to support 8 GPU cards
8. **Power Supply**
   4 x 2000W (2+2) Platinum Level efficiency redundant power supply

PCIe Topologies

SYS-4029GP-TRT
Dual-Root Topology



SYS-4029GP-TRT2
Single-Root Topology



FROM	TO	SYS-4029GP-TRT (uSEC)	SYS-4029GP-TRT2 (uSEC)
GPU1	GPU2	6.6	6.6
GPU1	GPU4	6.7	6.6
GPU1	GPU8	21.2	6.7

X11 for DEEP LEARNING/AI

Best–in-class technology designed for augmented for fast Deep Learning Training

10 4029GP-TRT2

Flexibility

- DP SKYLAKE CPU; 3UPI
- 24 DDR4 DIMMs
- 24 HS NVMe HDD bays
- 10 Double-Wide devices
- 12 x16 PCIe 3.0 slot
- 4 (2+2) 2000W Titanium PWS

4 1029GQ-TVRT

Scalability

- DP SKYLAKE CPU; 3UPI
- 12 DDR4 DIMMs
- 2 HS HDD bays
- 4 SXM w/ NVLink
- 4 x16 PCIe 3.0 slot
- 2 2000W Titanium PWS

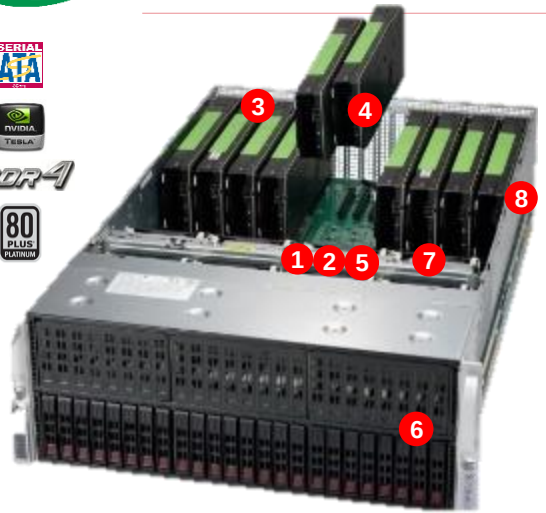
8 4029GP-TVRT

HyperScale

- DP SKYLAKE CPU; 3UPI
- 24 DDR4 DIMMs
- 16 HS HDD bays (w/ NVMe)
- 8 Pascal w/ NVLink
- 6 x16 PCIe 3.0 slot
- 4 (2+2) 2000W Titanium PWS

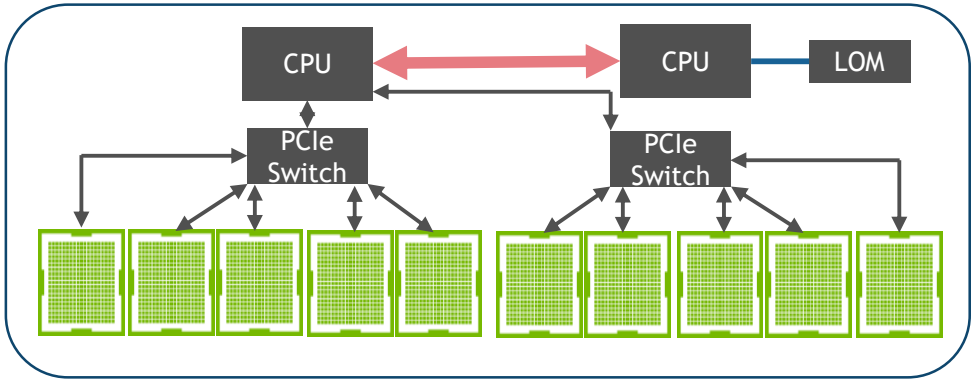
Support Volta-SXM2 Form Factor GPUs with Next Gen NVLink

SYS-4029GP-TRT2



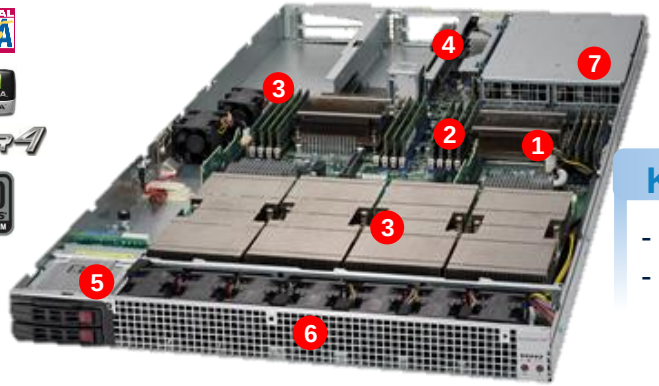
Key Features:

- 10 x16 PCIe3 GPUs under a single PCIe Root Complex
- Supports GPUDirect RDMA
- Supports up to 205W CPUs



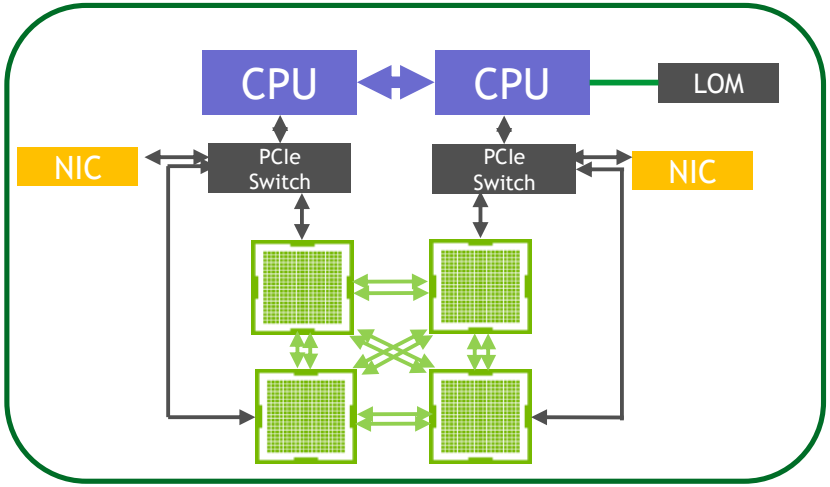
1. **Processor Support**
   Dual Xeon Scalable Processor; 3 UPI
2. **Memory Capacity**
   24 DIMMs ECC DDR4 2666 MHz
3. **Expansion Slots**
   11 PCI-e 3.0 x16 (10 double-wide slots for GPU)
   1 PCI-e3.0 x8
4. **I/O ports**
   1x VGA, 2x 10GbaseT LAN, 4x USB 2.0, and 1x IPMI dedicated LAN port, 1x M.2 NVMe
5. **System management**
   On board BMC (Baseboard Management Controllers) supports IPMI2.0, media/KVM over LAN with dedicated LAN for system management
6. **Drive Bays**
   24 hot-swap 2.5” NVMe drives bay
7. **System Cooling**
   8 heavy duty fans optimize to support 8 GPU cards
8. **Power Supply**
   4 x 2000W (2+2) Titanium Level efficiency redundant power supply

SYS-1029GQ-TVRT



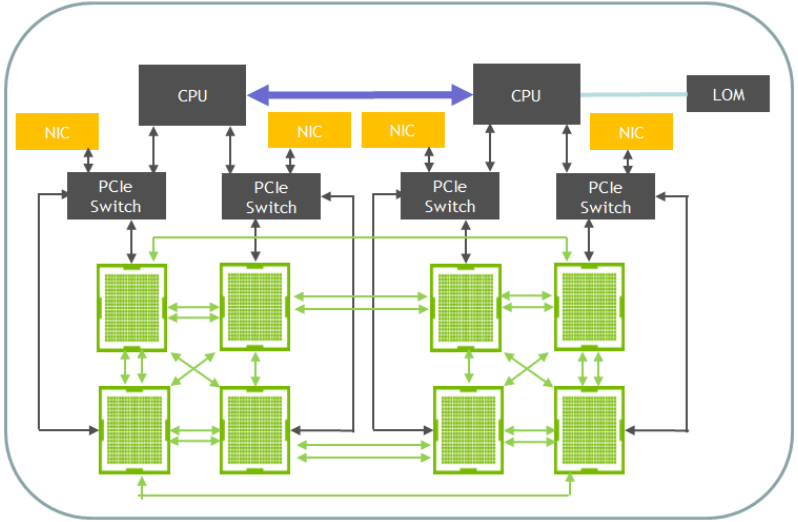
Key Features:

- NVIDIA Tesla V100 Enabled
- Optimized for GPUDirect RDMA



1. **Processor Support**
   Dual Xeon Scalable Processor; 3 UPI
   Quad Tesla V100 SXM2 GPUs
2. **Memory Capacity**
   12 DIMMs ECC DDR4 2666 MHz
3. **Expansion Slots**
   2 x16 (FHFL/LP) from PLX;
   2 x16 (FHFL/LP) from CPU
4. **I/O ports**
   1x VGA, 2x 10GbaseT LAN, 2x USB 3.0, and 1x IPMI dedicated LAN port
5. **Drive Bays**
   2x HS 2.5” NVMe drives bays; 4x total 2.5” HDD bays
6. **System Cooling**
   7 counter rotating fans with optimal fan speed
7. **Power Supply**
   2000W Titanium redundant power supply

SYS-4029GP-TVRT



Processor Support

Dual Xeon Scalable Processor; 3 UPI

8 Tesla SXM2 V100 GPUs

Memory Capacity

24 DIMMs ECC DDR4 2666 MHz

Expansion Slots

4 PCI-e 3.0 x16 LP (via RDMA for IB EDR)

2 PCI-e 3.0 x16 LP

I/O ports

1x VGA, 2x 10G-BaseT LAN, 3x USB 3.0, and 1x IPMI dedicated LAN port, 1x M.2 NVMe

Drive Bays

16 hot-swap 2.5” drives bay (Support up to 8x NVMe)

System Cooling

8 heavy duty fans optimize to support 8 GPU cards

Power Supply

4 x 2200W (2+2) Titanium Level efficiency redundant power supply

New X11 Products

New Systems developed for best in class Deep Learning Inference and Training

20 6049GP-TRT

Inference & Transcoding

- DP SKYLAKE CPU; 3UPI
- 20 Single-Wide GPUs
- 24 DDR4 DIMMs
- 24 3.5” HDD bays
- 4 (2+2) 2000W Titanium PWS

16 9029GP-TNVRT

DL Training

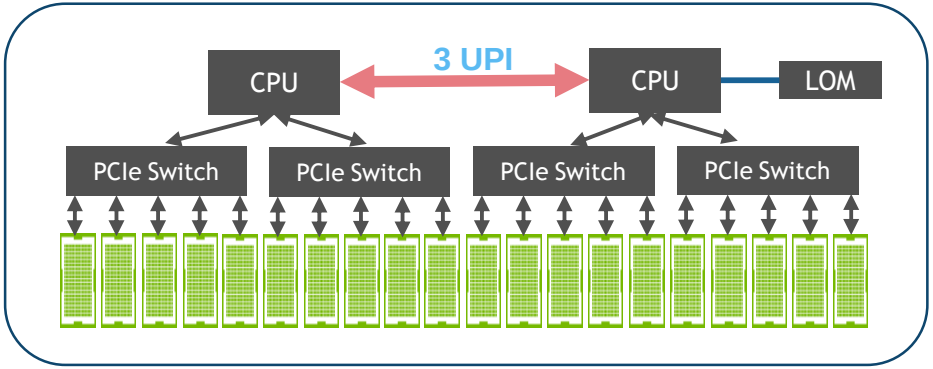
- DP SKYLAKE CPU; 3UPI
- 16 SXM3GPU
- NVSwitch & NVLink
- 24 DDR4 DIMMs
- 16 NVMe U.2 drive bays
- 16 PCIe x16 for RDMA
- 6 3000W Titanium PWS

20 GPU System: SYS-6049GP-TRT

Key Features:

- 20 single width GPUs
- Dual Xeon Scalable Processor; 205W
- 24 hot swappable 3.5” drives

328 Lanes of PCIe



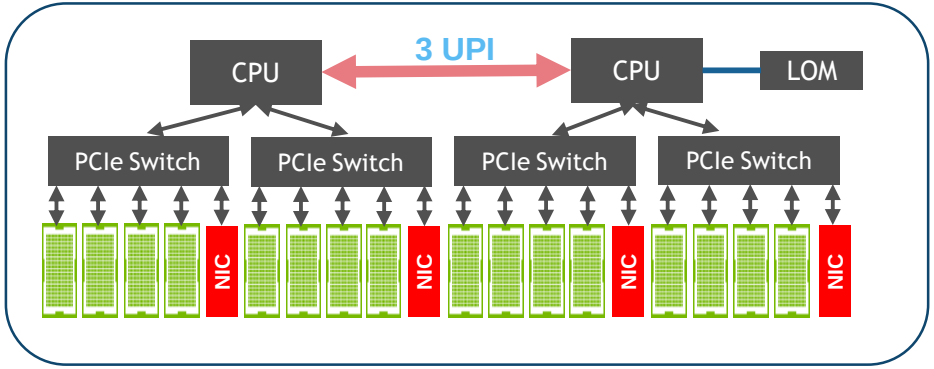
1. **Processor Support**
   Dual Xeon Scalable Processor; 3 UPI
2. **Memory Capacity**
   24 DIMMs ECC DDR4 2666 MHz
3. **Expansion Slots**
   20 PCIe 3.0 x16 for single-wide GPU cards
   1 PCIe 3.0 x8 (FHFL x16 slots)
4. **I/O ports**
   1x VGA, 2x 10GbaseT LAN, 4x USB 3.0, and 1x IPMI dedicated LAN port, 1x M.2 NVMe
5. **System management**
   On board BMC (Baseboard Management Controllers) supports IPMI2.0
6. **Drive Bays**
   24 hot-swap 3.5” drive bays
7. **System Cooling**
   8 heavy duty fans optimize to support 8 GPU cards
8. **Power Supply**
   4 x 2000W (2+2) Titanium Level efficiency redundant power supply

20 GPU System: SYS-6049GP-TRT

Key Features:

- 20 single width GPUs
- Dual Xeon Scalable Processor; 205W
- 24 hot swappable 3.5” drives

328 Lanes of PCIe



1. **Processor Support**
   Dual Xeon Scalable Processor; 3 UPI
2. **Memory Capacity**
   24 DIMMs ECC DDR4 2666 MHz
3. **Expansion Slots**
   20 PCIe 3.0 x16 for single-wide GPU cards
   1 PCIe 3.0 x8 (FHFL x16 slots)
4. **I/O ports**
   1x VGA, 2x 10GbaseT LAN, 4x USB 3.0, and 1x IPMI dedicated LAN port, 1x M.2 NVMe
5. **System management**
   On board BMC (Baseboard Management Controllers) supports IPMI2.0
6. **Drive Bays**
   24 hot-swap 3.5” drive bays
7. **System Cooling**
   8 heavy duty fans optimize to support 8 GPU cards
8. **Power Supply**
   4 x 2000W (2+2) Titanium Level efficiency redundant power supply

Dawn of a New Age of AI

- **Dawn of a new age of AI driven by Deep Learning**
  - Natural speech, Autonomous Mobility, Medical Image based diagnosis many others
- **AI models continue to increase in size, requiring weeks to train**
  - Google 'Mixture of Experts' has 8 Billion parameters (up from 100M, 2 years ago)
- **Supermicro HGX-2 system is a powerful Deep Learning System**
  - 16 V100 32G GPUs powered by NVLink & NVSwitch
  - High throughput, low latency interconnect between GPUs
  - Up to 2.7X faster training
  - 2 petaFLOPs of AI performance

- **Versatile System for Cloud Service Providers**
  - Hypervisor based option to virtualize number of GPUs (1, 2, 4, 8, 16) for target workload

SYS-9029GP-TNVRT

10U System
Includes CPU head node

- NVLink + NVSwitch based high performance GPU Interconnect

Processor Support
Dual Xeon Scalable Processor; 3 UPI
16 Tesla V100 32GB SXM3 GPUs

Memory Capacity
24 DIMMs ECC DDR4 2666 MHz

Expansion Slots
16 PCI-e 3.0 x16 LP (via RDMA for IB EDR)
2 PCI-e 3.0 x16 LP

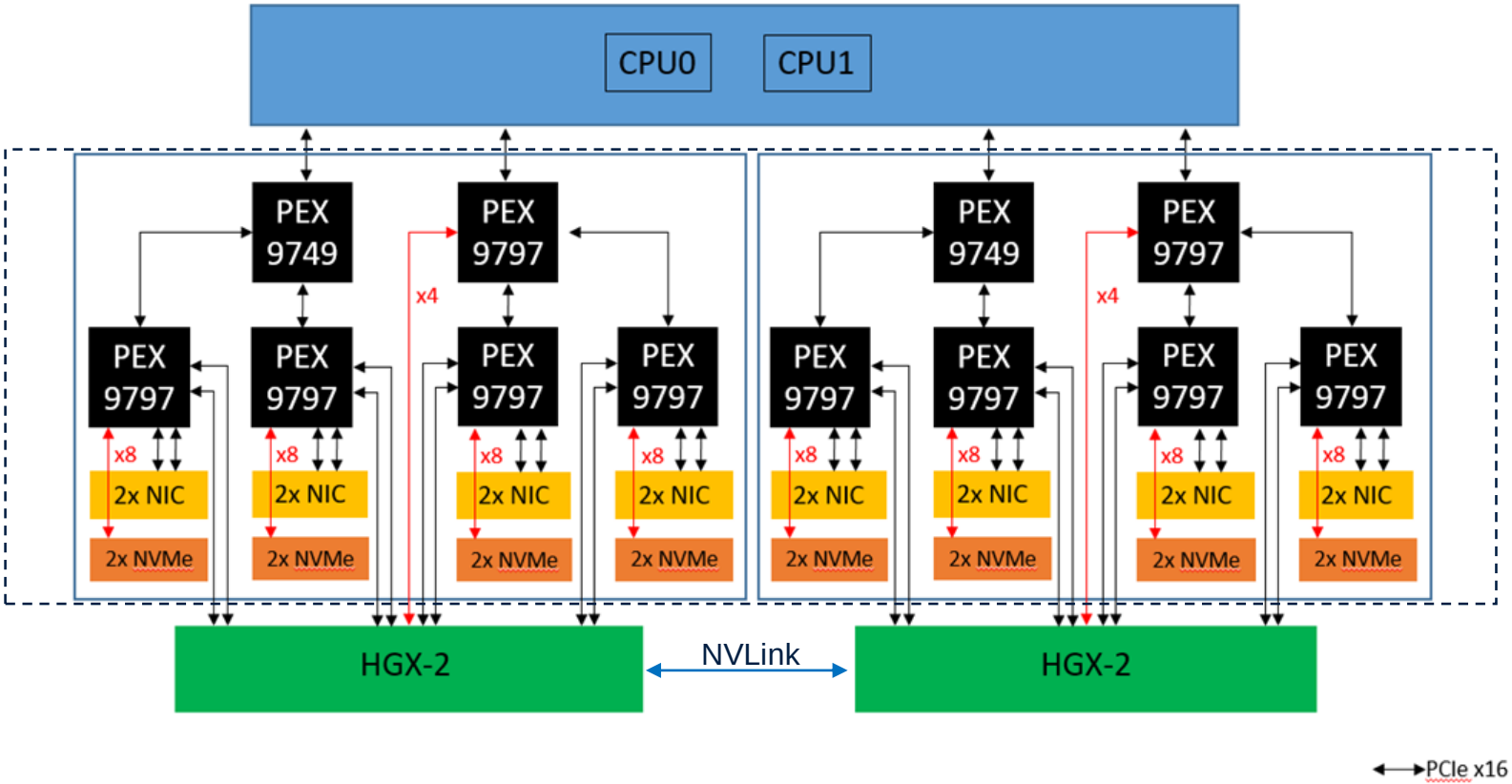
I/O ports
1x VGA, 2x 10G-BaseT LAN, 3x USB 3.0, and 1x IPMI dedicated LAN port

Drives
16 NVMe U.2 2.5” drives bays & 6 SATA 2.5” drives bays
2 M.2 NVMe

System Cooling
14 heavy duty fans

Power Supply
6 x 3000WTitanium Level efficiency power supplies

9029GP - PCIe Switch Module Block Diagram



PCIe Switch Module

Top to Bottom V100 NVLink Systems

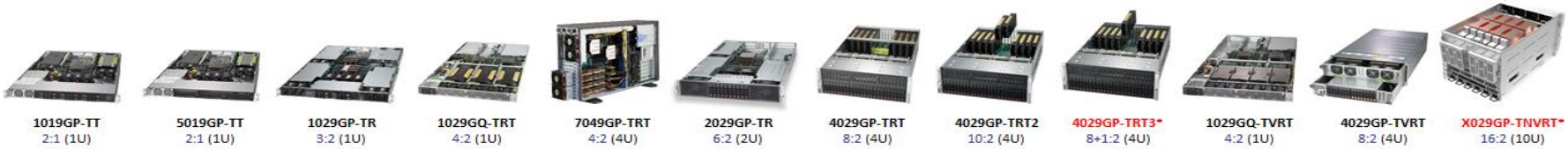
	SYS-1029GP-TVRT	SYS-4029GP-TVRT	SYS-9029GP-TNVRT
GPU	4 NVIDIA Tesla V100 (SXM2)	8 NVIDIA Tesla V100 (SXM2)	16 NVIDIA Tesla V100 (SXM3)
Performance	0.5 PetaFLOPS	1 PetaFLOPS	2 PetaFLOPS
CUDA Cores	20,480	40,960	81,920
Tensor Cores	2,560	5,120	10,240
CPU : GPU	2 : 4	2 : 8	2 : 16

Retrofitting existing Compute Infrastructure

- Happens in all industry when faced with paradigm shifts or pushing the limits of existing technology
  - Electric Cars (Paradigm shift – moving from Internal Combustion Engine to Electric Motor)
    - Cannot leverage existing worldwide infrastructure of gas stations
    - Need a full new ecosystem – charging stations; cannot happen over night
  - Airbus A380 (Pushing the limits – complete double decker aircraft)
    - Requires updated airport infrastructure
    - Wider taxiways, update Jet bridges, Reinofrced runways
- Compute Industry also needs to update to support higher power processors
  - Power Delivery
    - Rack Power levels need to increase, traditional CPU budgets were 75 to 130W, but now over 200W and going higher
    - GPUs are at 300 to 350W, single 1U node can have 2 CPUs and 4 GPUs
    - Need higher levels of power distribution in the rack and the data center
  - Power Dissipation
    - In some systems fans are at the limit of noise safety limits
  - Liquid Cooling starting to become more common in some datacenters

Conclusion

- Moore's Law has been one heck of a ride, but the sun is setting on it
- AI is the killer app today
  - Disruptive in many industries, will impact multiple walks of life
- Both the above points are driving significant change and innovation in the computer industry
- GPU is the parallel compute engine of today
  - We see an ever increasing number of applications getting 'GPU Accelerated' (eg databases); its much more than graphics
- Supermicro offers one of the widest range of GPU systems in the industry



Thank You