Dell Presentation Template Standard 4:3 Layout

Presenter Name

Title

Marcel van Drunen – Dell
Enterprise Technologist HPC

- Short intro Dell
- Dell and High Performance Computing
- The HPC market
- When a workstation is not enough
- GPGPU
- Cluster
- (virtual) SMP
- Components Infiniband/10GigE/Filesystems(Panasas)
- Dell portfolio?/Terminology (grid, cloud, hpc, etc)/Intel&AMD

“We’re focused on scalable and flexible solutions that simplify high-performance computing by reducing cost and complexity.

What we’re learning about HPC technology will redefine productivity throughout the research, discovery and business computing ecosystem.”

Michael Dell - 2008

Definitions

- High Performance Computing (HPC)
  - Computing aimed at calculations, not at transactions.
- High Performance Compute Cluster (HPCC)
  - Set of computers that provide compute power, not redundancy.
- Grid
  - Geographically dispersed set of (compute) resources.
- (Compute) Cloud
  - Scalable pool of (compute) resources that hides complexity form users and management, pay-per-use model

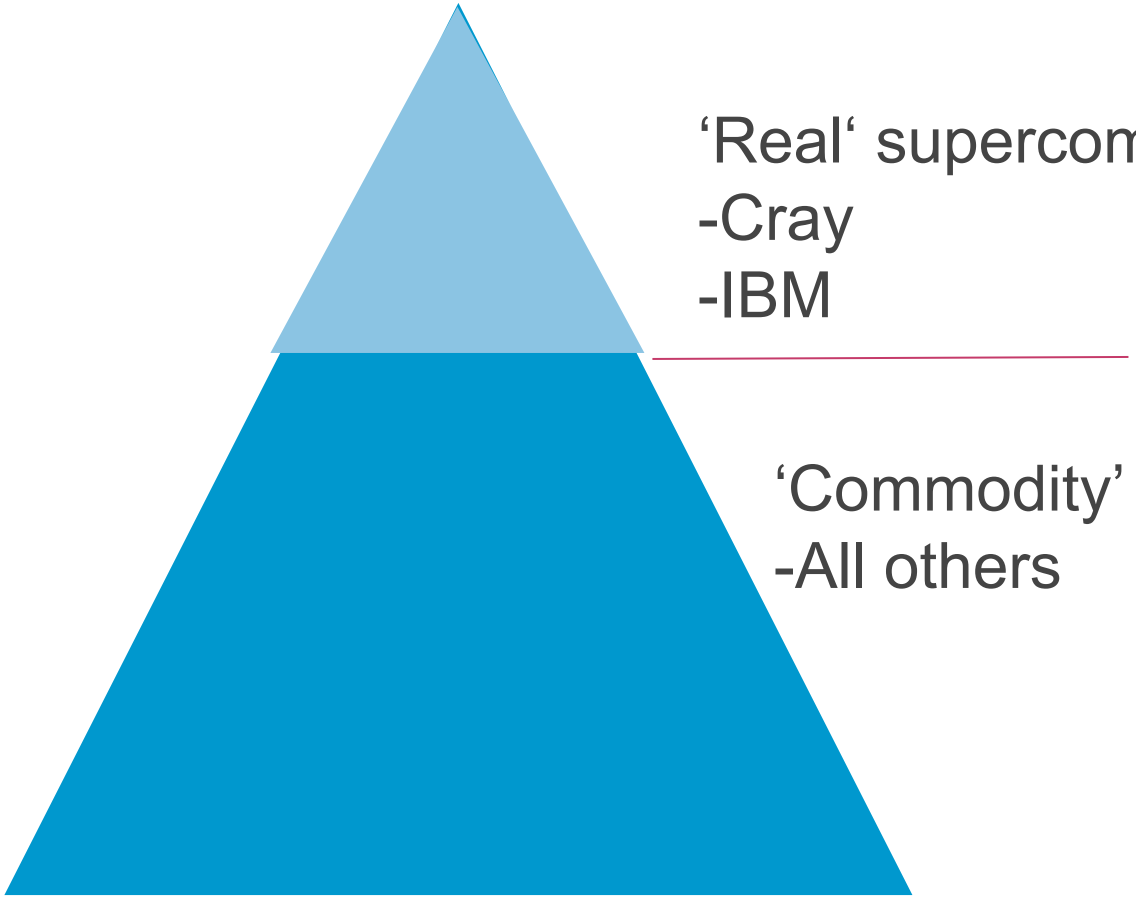
Dell's mission in the HPC world

- Dell uses commodity, best of breed components to **simplify** HPC by driving out **cost** and **complexity**. This makes HPC available to a larger amount of researchers.

- Dell has done the same to other markets:
  - Desktops
  - Laptops
  - Servers
  - Storage

The High Performance Computing market

Market
value:
10 B$/year

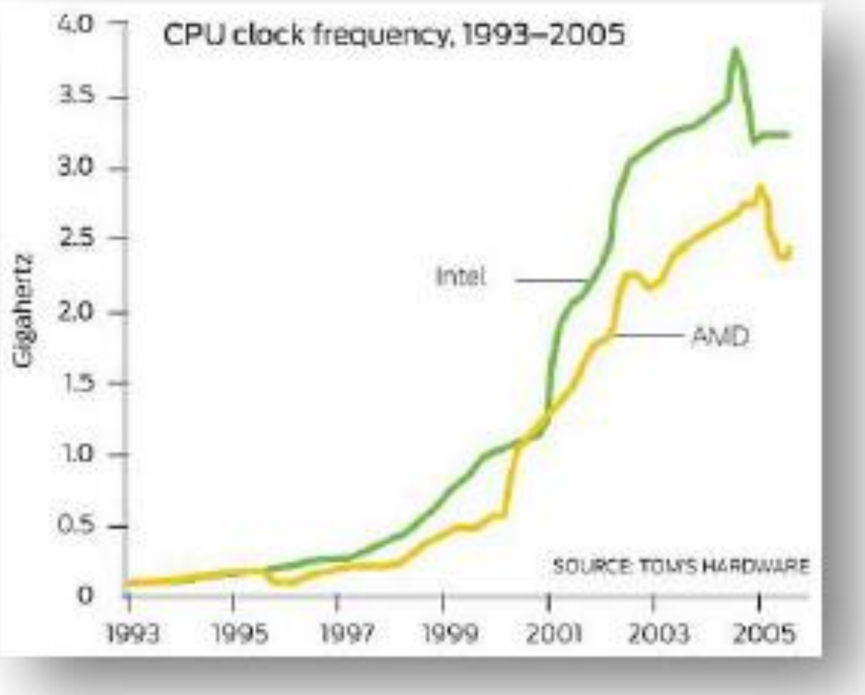


‘Real‘ supercomputing
-Cray
-IBM

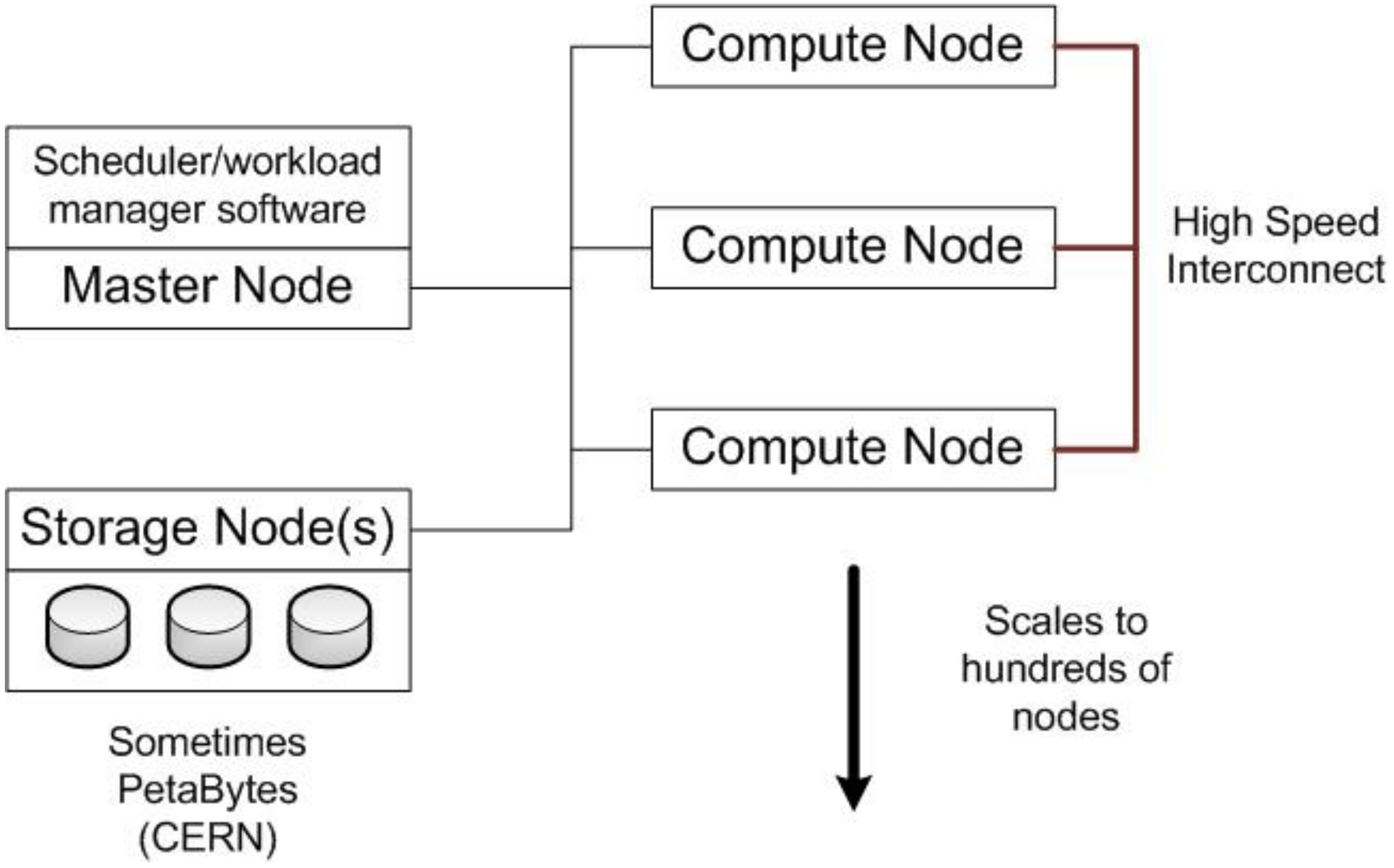
‘Commodity’ computing
-All others

When a workstation is not enough

- Server CPU’s are not faster than workstation CPU’s
- Parallelize code!
  - MPI
  - OpenMP
  - CUDA/OpenCL
- Hardware choices
  - Cluster (HPCC)
  - (Virtual)SMP machine
  - GPGPU



Typical HPC Cluster



Scheduler/workload manager software

Master Node

Storage Node(s)

Sometimes PetaBytes (CERN)

Compute Node

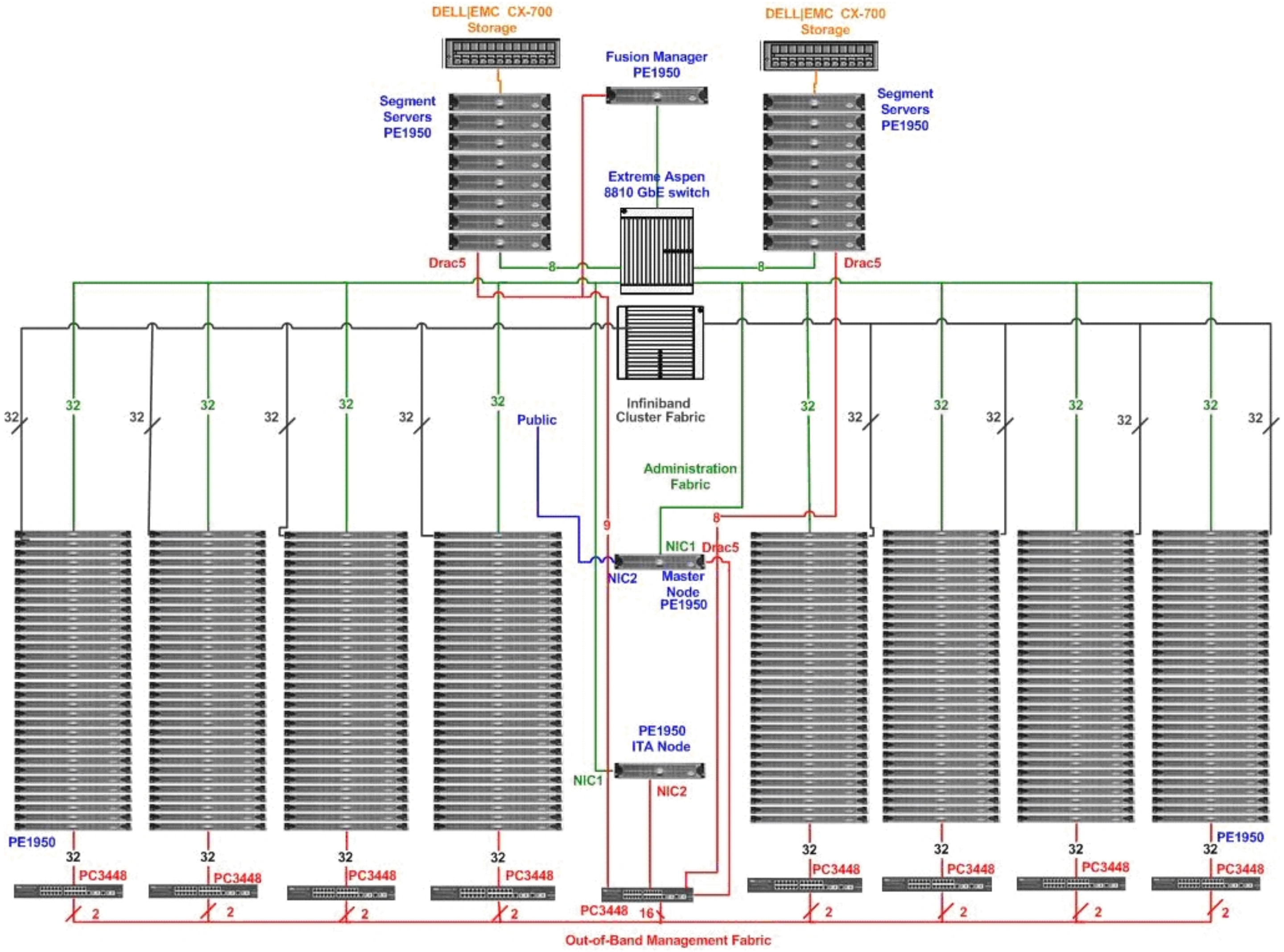
Compute Node

Compute Node

High Speed Interconnect

Scales to hundreds of nodes

256 Node Cluster (3072 cores/12 TB memory)



Interconnects

- Infiniband is de-facto standard
  - QDR Infiniband 40Gb/s, 80/160 Gb/s under development
  - Very low latency (microsecond)

- 10GigE Ethernet is gaining marketshare
  - 10 Gb/s, 40/100 Gb/s under development
  - Much improvement in latency (needs Fiber connection)

- Talk of the town: ‘converged’ networking

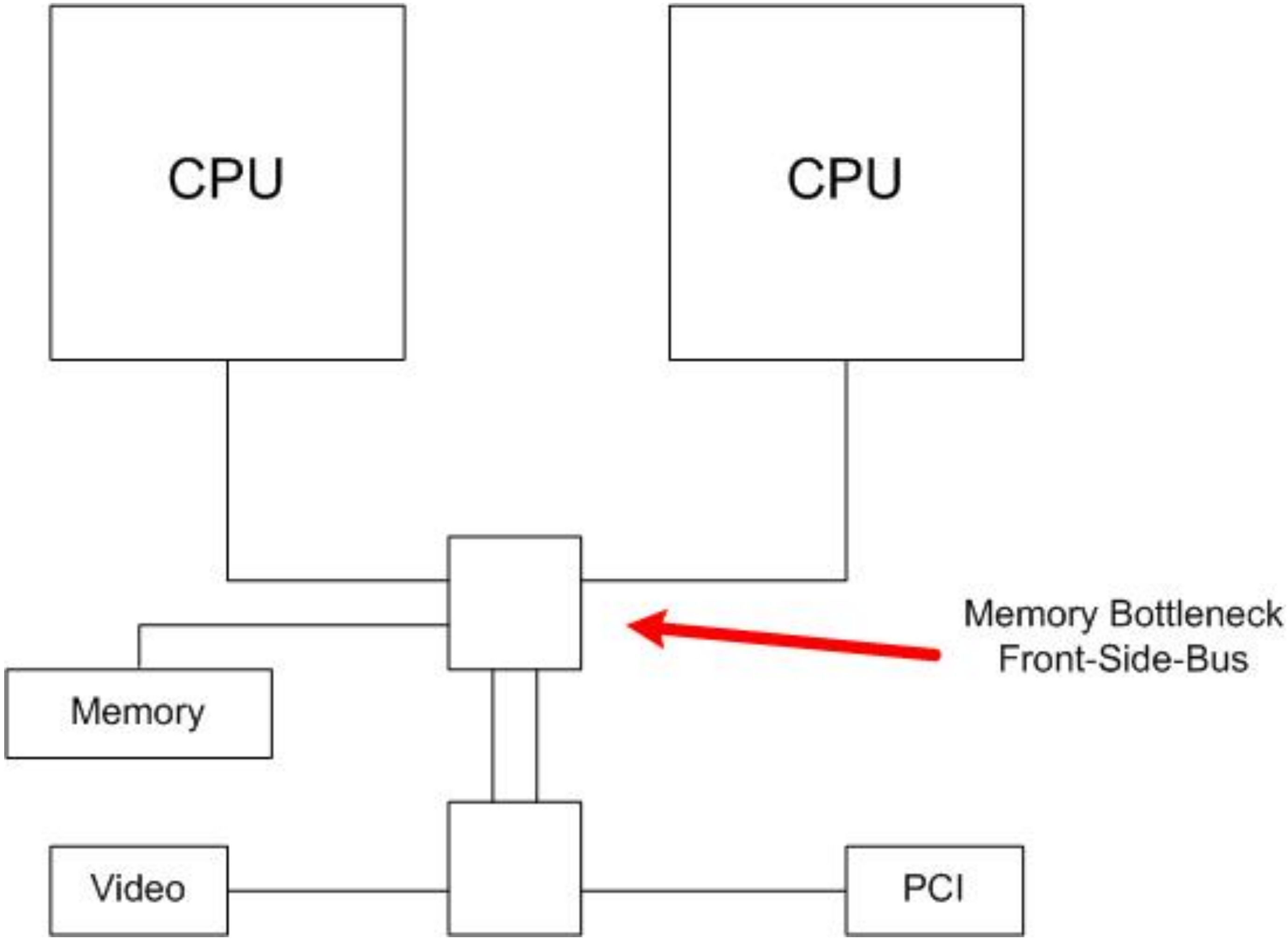
Mellanox TECHNOLOGIES

QLOGIC The Ultimate in Performance

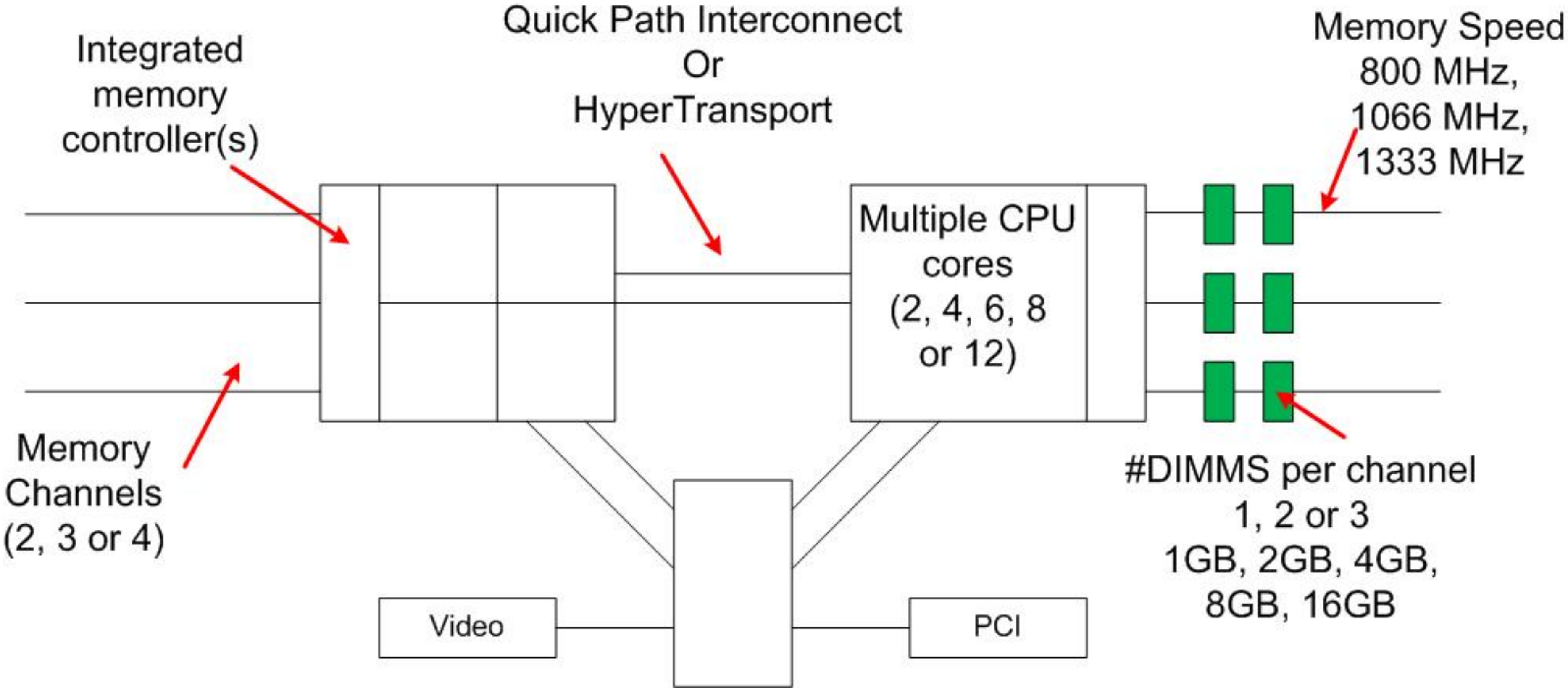
Developments in CPU design

- Clockspeed doesn’t rise anymore
- Power consumption is now an issue
- Dell uses Intel and AMD for 1, 2 and 4 socket machines
- For most uses, CPU is not bottleneck (NON-HPC)
- Memory bandwidth becomes very important
- Commodity CPU should support Virtualization, Security
- Smaller process allow for integration of non-CPU components
  - Memory controller
  - GPU, PCI, RAID-controller, etc. etc.

Old Intel & AMD infrastructure



Current AMD/Intel Architecture



Integrated memory controller(s)

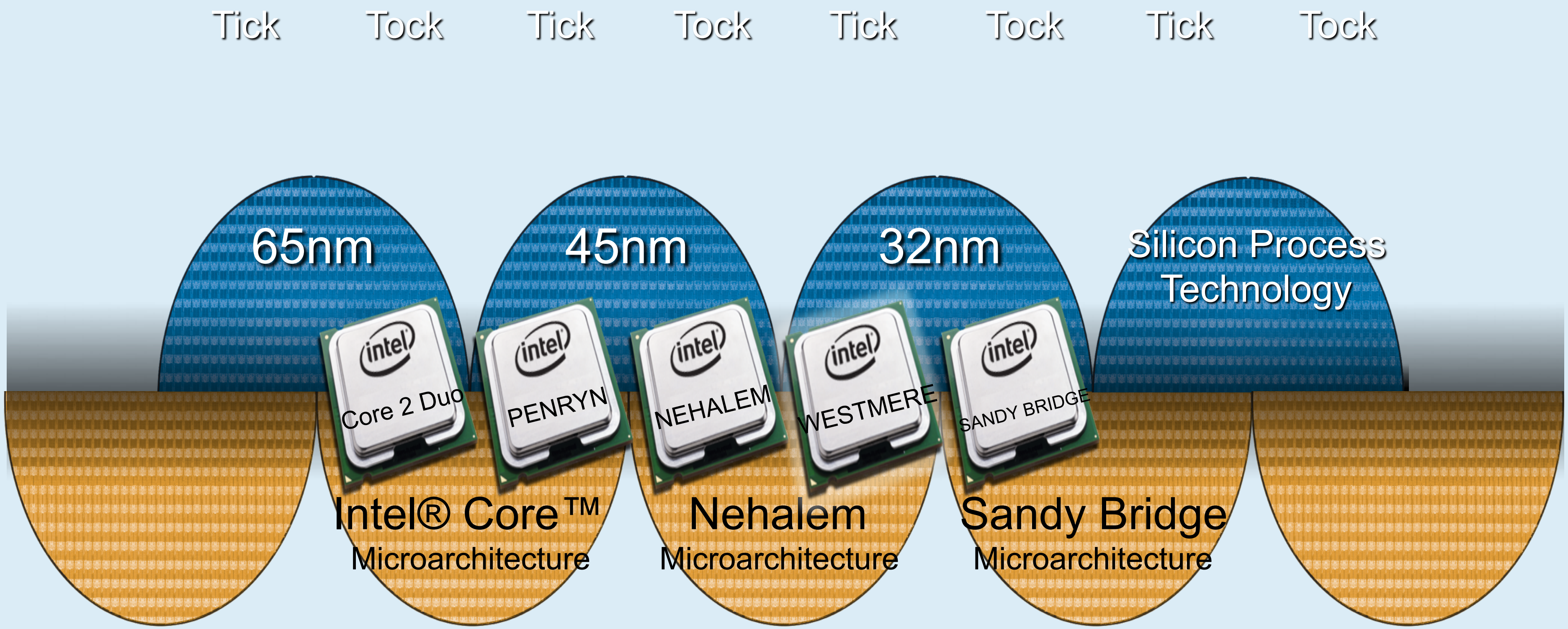
Quick Path Interconnect Or HyperTransport

Memory Speed 800 MHz, 1066 MHz, 1333 MHz

Memory Channels (2, 3 or 4)

#DIMMS per channel 1, 2 or 3 1GB, 2GB, 4GB, 8GB, 16GB

Maintaining the Pace of Execution: Tick-Tock Microprocessor Development



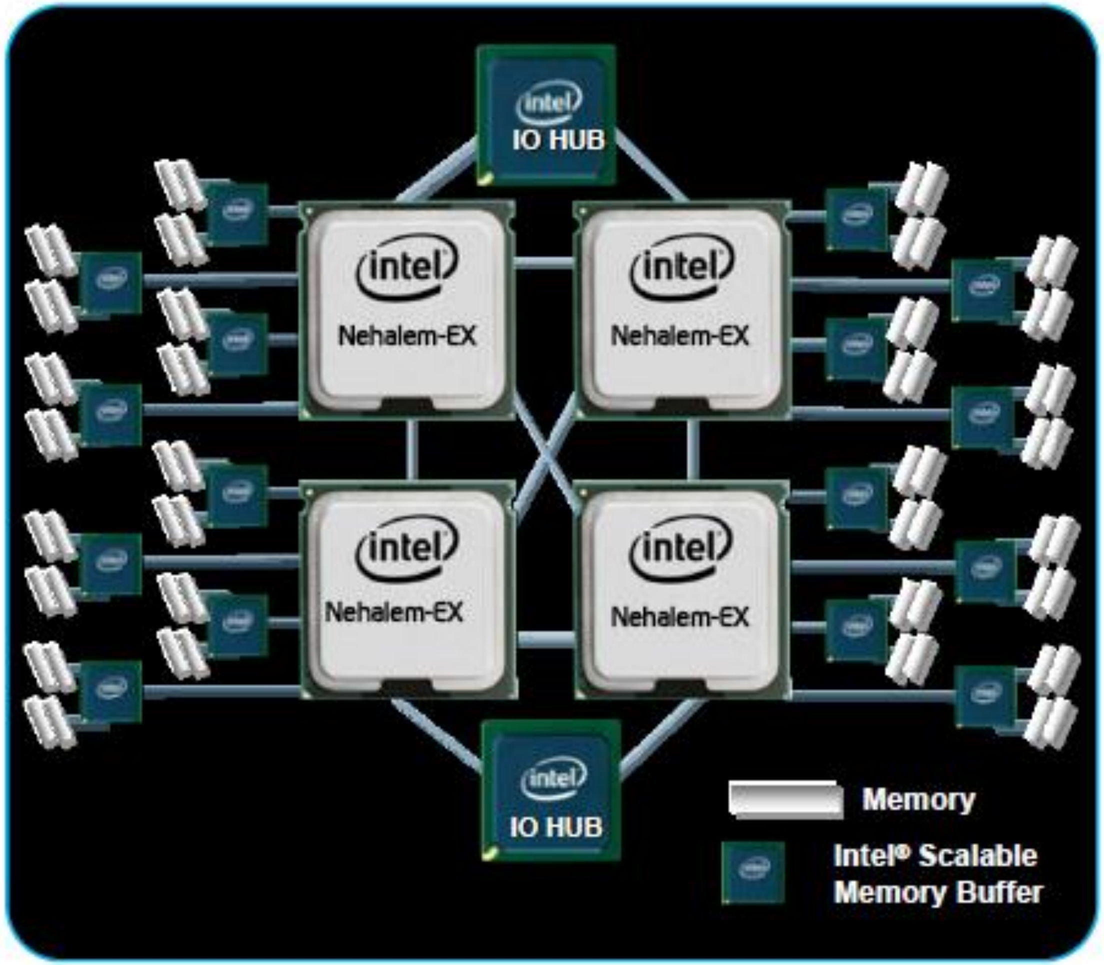
32nm Process Health Enables Acceleration of Westmere Product Ramp

Intel (launched in March’10

- Westmere EP (XEON 5600 series)
  - 2 socket
  - 32nm (Die-shrink of Nehalem EP)
  - 3 memory lanes per CPU, up to 1333 MHz
  - Up to 6 core

- Nehalem EX (XEON 7500 series)
  - Up to 8 socket (Dell up to 4 socket)
  - 45nm (expect Westmere version end of this year)
  - 4 memory lanes per CPU
  - Up to 8 core

Nehalem-EX



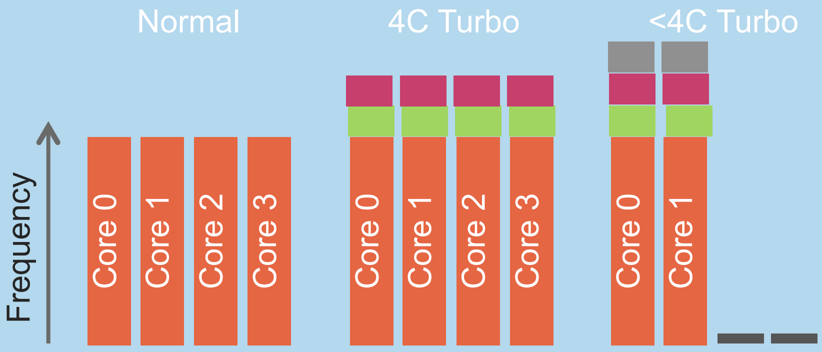
Up to 64 DIMM slots for up to (64 x 16GB) = 1 TerraByte of memory

Performance Enhancements

Intel Xeon® 5500/5600/6500/7500 Series Processor

Intel® Turbo Boost Technology

Increases performance by increasing processor frequency and enabling faster speeds when conditions allow



All cores operate at rated frequency

All cores operate at higher frequency

Fewer cores may operate at even higher frequencies

Higher performance on demand

Intel® Hyper-Threading Technology

Increases performance for threaded applications delivering greater throughput and responsiveness

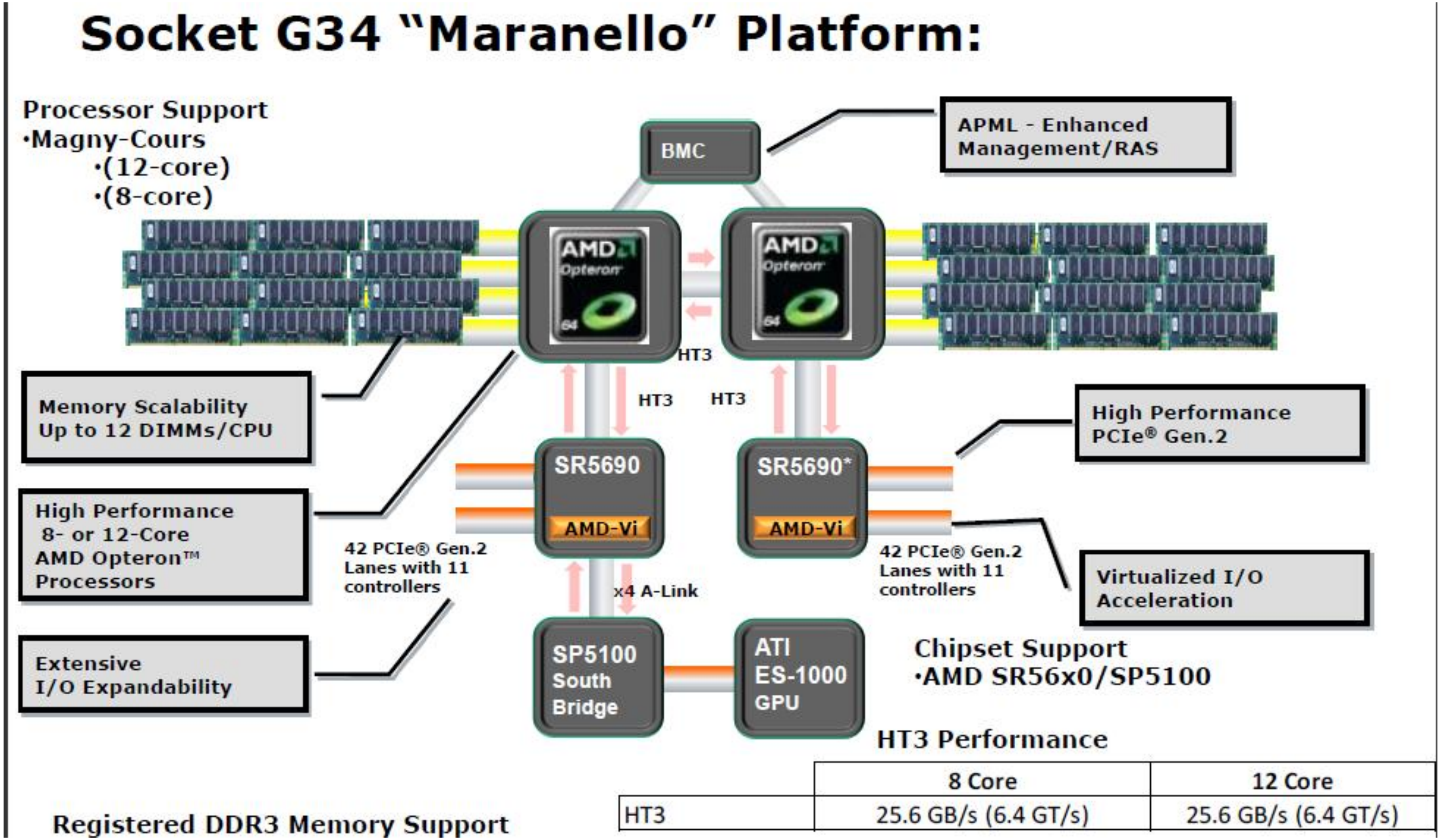
Thread 1
Thread 2
Thread 3
Thread 4
Thread 5
Thread 6
Thread 7
Thread 8
Core
Core
Core
Core
Nehalem Microarchitecture

Up to 30% higher†

Higher performance for threaded workloads

† For notes and disclaimers, see performance and legal information slides at end of this presentation.

AMD 2/4 socket Maranello



Socket G34 "Maranello" Platform:

Processor Support
- Magny-Cours
  - (12-core)
  - (8-core)

APML - Enhanced Management/RAS

Memory Scalability Up to 12 DIMMs/CPU

High Performance 8- or 12-Core AMD Opteron™ Processors

Extensive I/O Expandability

High Performance PCIe® Gen.2

Virtualized I/O Acceleration

42 PCIe® Gen.2 Lanes with 11 controllers

42 PCIe® Gen.2 Lanes with 11 controllers

Chipset Support
- AMD SR56x0/SP5100

HT3 Performance

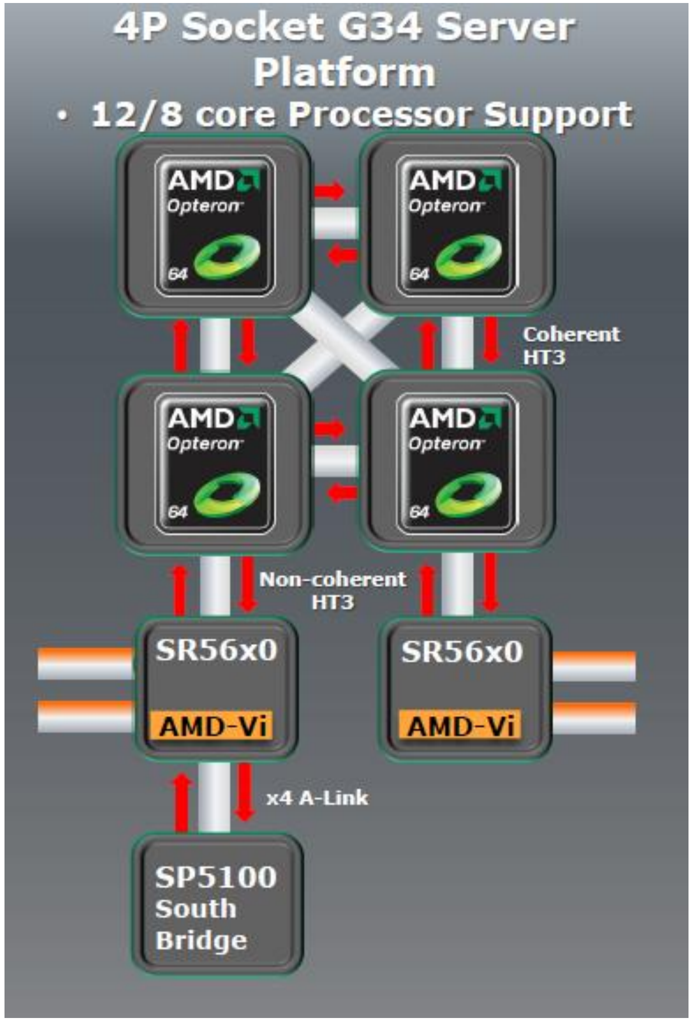
	8 Core	12 Core
HT3	25.6 GB/s (6.4 GT/s)	25.6 GB/s (6.4 GT/s)

Registered DDR3 Memory Support

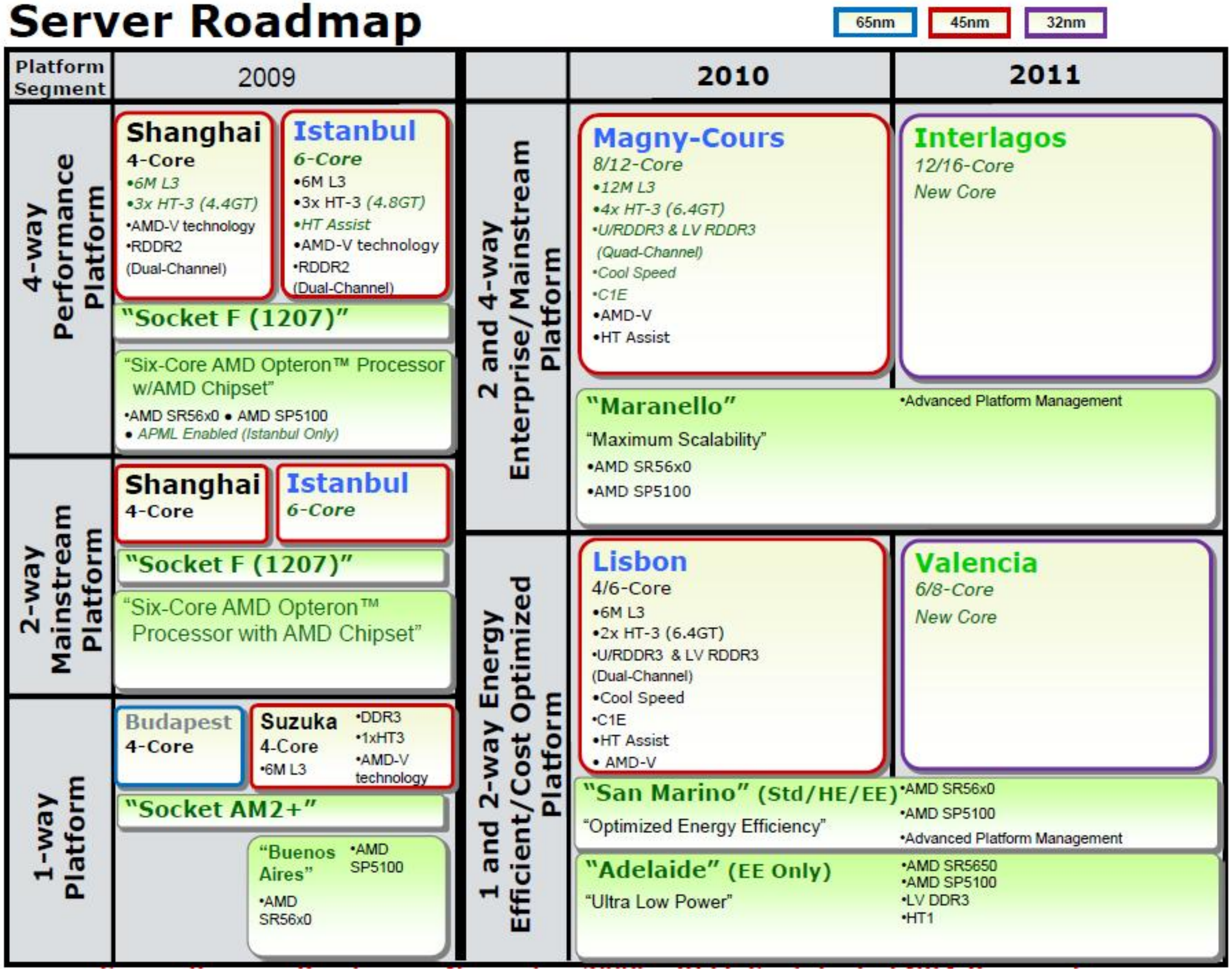
AMD 4 socket

-4 x 12 cores = 48 cores!

-Dell R815 can contain up to 512GB of memory



Server Roadmap



65nm 45nm 32nm

Platform Segment | 2009

4-way Performance Platform

Shanghai
4-Core
- 6M L3
- 3x HT-3 (4.4GT)
- AMD-V technology
- RDDR2 (Dual-Channel)

Istanbul
6-Core
- 6M L3
- 3x HT-3 (4.8GT)
- HT Assist
- AMD-V technology
- RDDR2 (Dual-Channel)

"Socket F (1207)"

"Six-Core AMD Opteron™ Processor w/AMD Chipset"
- AMD SR56x0 ● AMD SP5100
- APML Enabled (Istanbul Only)

2-way Mainstream Platform

Shanghai
4-Core

Istanbul
6-Core

"Socket F (1207)"

"Six-Core AMD Opteron™ Processor with AMD Chipset"

1-way Platform

Budapest
4-Core

Suzuka
4-Core
- 6M L3
- DDR3
- 1xHT3
- AMD-V technology

"Socket AM2+"

"Buenos Aires"
- AMD SP5100
- AMD SR56x0

2010 | 2011

2 and 4-way Enterprise/Mainstream Platform

Magny-Cours
8/12-Core
- 12M L3
- 4x HT-3 (6.4GT)
- U/RDDR3 & LV RDDR3 (Quad-Channel)
- Cool Speed
- C1E
- AMD-V
- HT Assist

Interlagos
12/16-Core
New Core

"Maranello"
"Maximum Scalability"
- AMD SR56x0
- AMD SP5100
- Advanced Platform Management

1 and 2-way Energy Efficient/Cost Optimized Platform

Lisbon
4/6-Core
- 6M L3
- 2x HT-3 (6.4GT)
- U/RDDR3 & LV RDDR3 (Dual-Channel)
- Cool Speed
- C1E
- HT Assist
- AMD-V

Valencia
6/8-Core
New Core

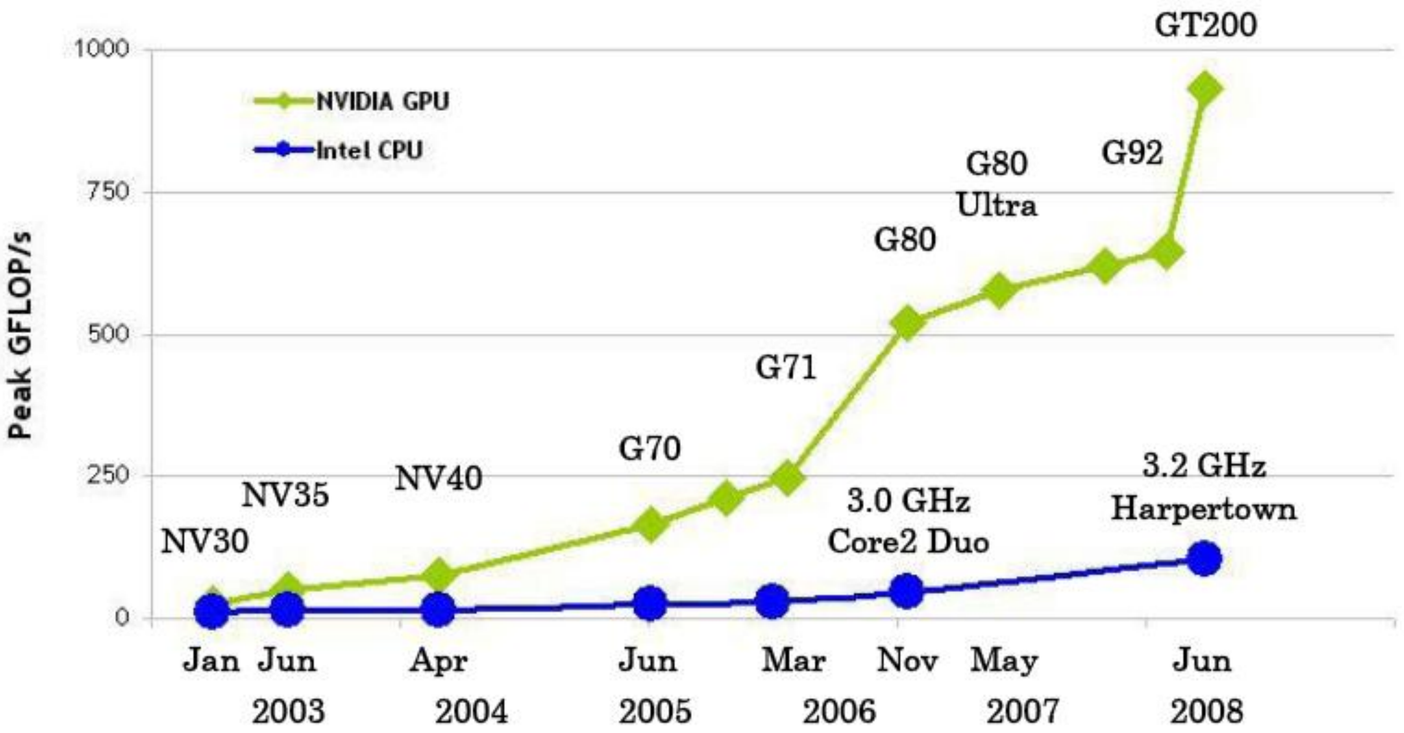
"San Marino" (Std/HE/EE)
"Optimized Energy Efficiency"
- AMD SR56x0
- AMD SP5100
- Advanced Platform Management

"Adelaide" (EE Only)
"Ultra Low Power"
- AMD SR5650
- AMD SP5100
- LV DDR3
- HT1

What’s a GPU?

- High-end video card adapted for computation
- nVidia or AMD/ATi
- Programmable with CUDA or Open-CL

The GPU proposition (1)



GT200 = GeForce GTX 280

G92 = GeForce 9800 GTX

G80 = GeForce 8800 GTX

G71 = GeForce 7900 GTX

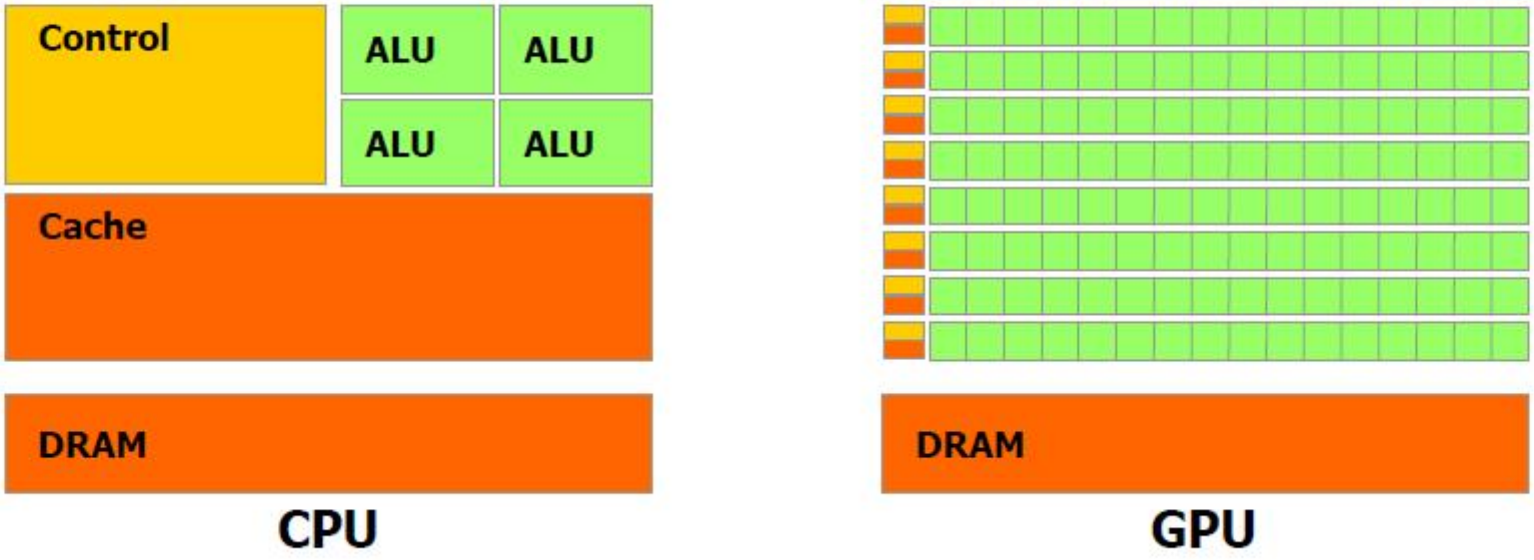
G70 = GeForce 7800 GTX

NV40 = GeForce 6800 Ultra

NV35 = GeForce FX 5950 Ultra

NV30 = GeForce FX 5800

The GPU proposition (2)



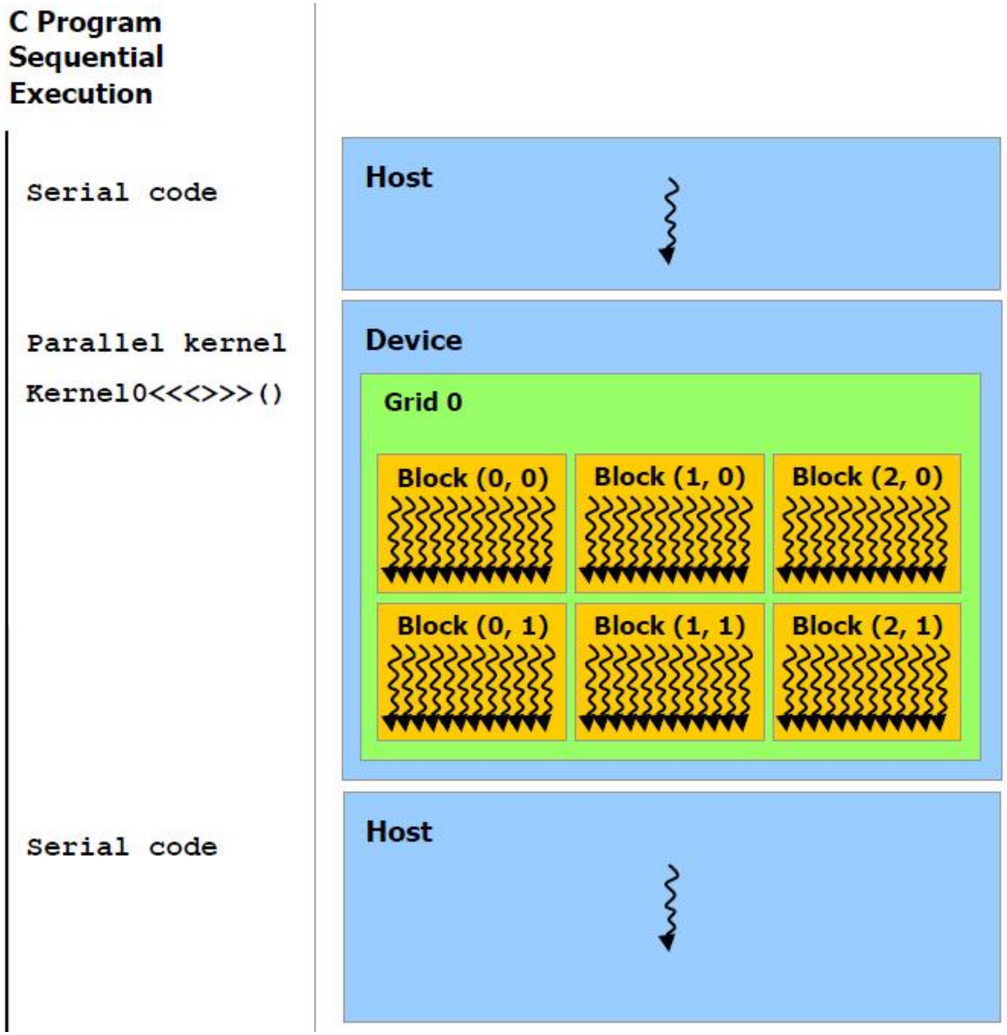
A CPU needs a lot of logical elements for all kinds of control functions. GPU’s are especially well-suited to address problems that can be expressed as data-parallel computations

CPU and GPU cooperating

-Some problems can be optimized for GPU

- Some will always run better on CPU

- The ideal machine has both



C Program Sequential Execution

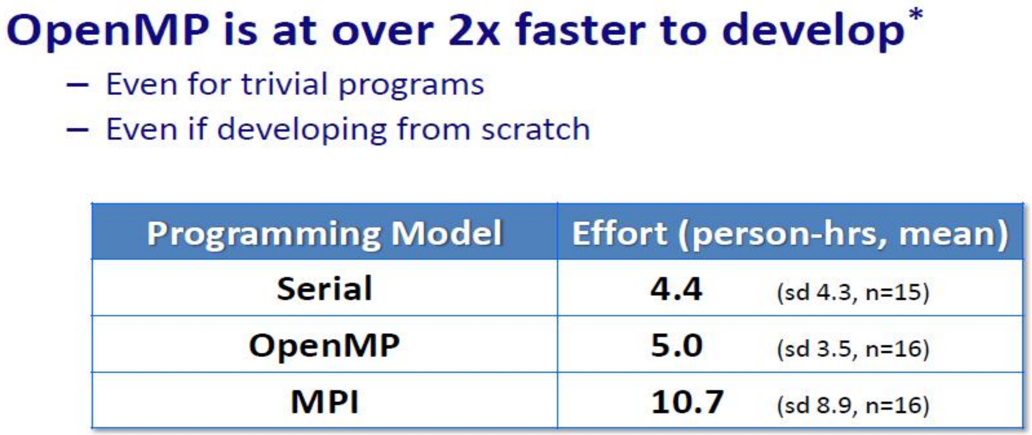
```
Serial code

Parallel kernel
Kernel0<<<>>>()

Serial code
```

What if 1TB is not enough?

- ‘Real big’ SMP machine, or:
- ScaleMP, virtual SMP machine
  - Can use OpenMP instead of MPI
  - Can scale with needs
  - Commodity hardware
    - Low cost
    - Easier maintenance



OpenMP is at over 2x faster to develop*

- Even for trivial programs
- Even if developing from scratch

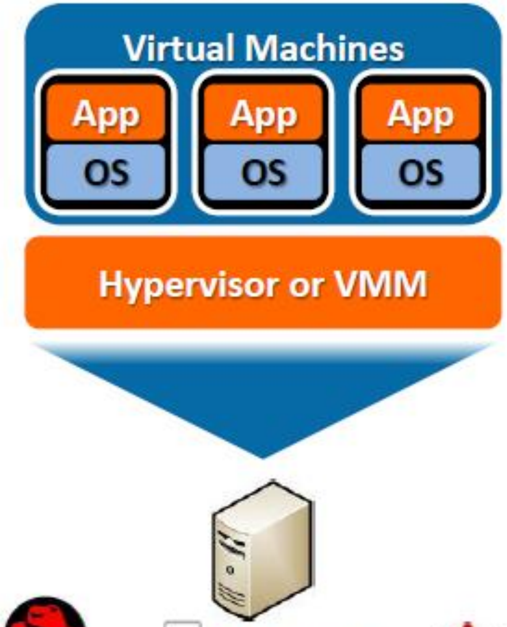
Programming Model	Effort (person-hrs, mean)
Serial	4.4 (sd 4.3, n=15)
OpenMP	5.0 (sd 3.5, n=16)
MPI	10.7 (sd 8.9, n=16)

The ScaleMP proposition:

PARTITIONING

Subset of a physical resource

(For applications requiring a fraction of the physical server resources)



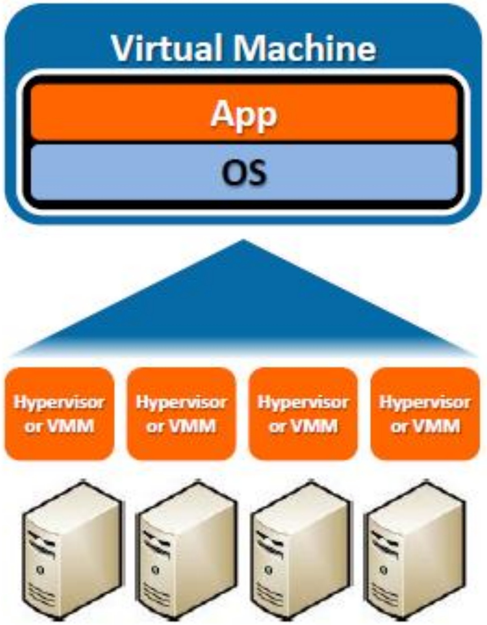
redhat vmware CITRIX

QUMRANET Microsoft Xen Source

AGGREGATION

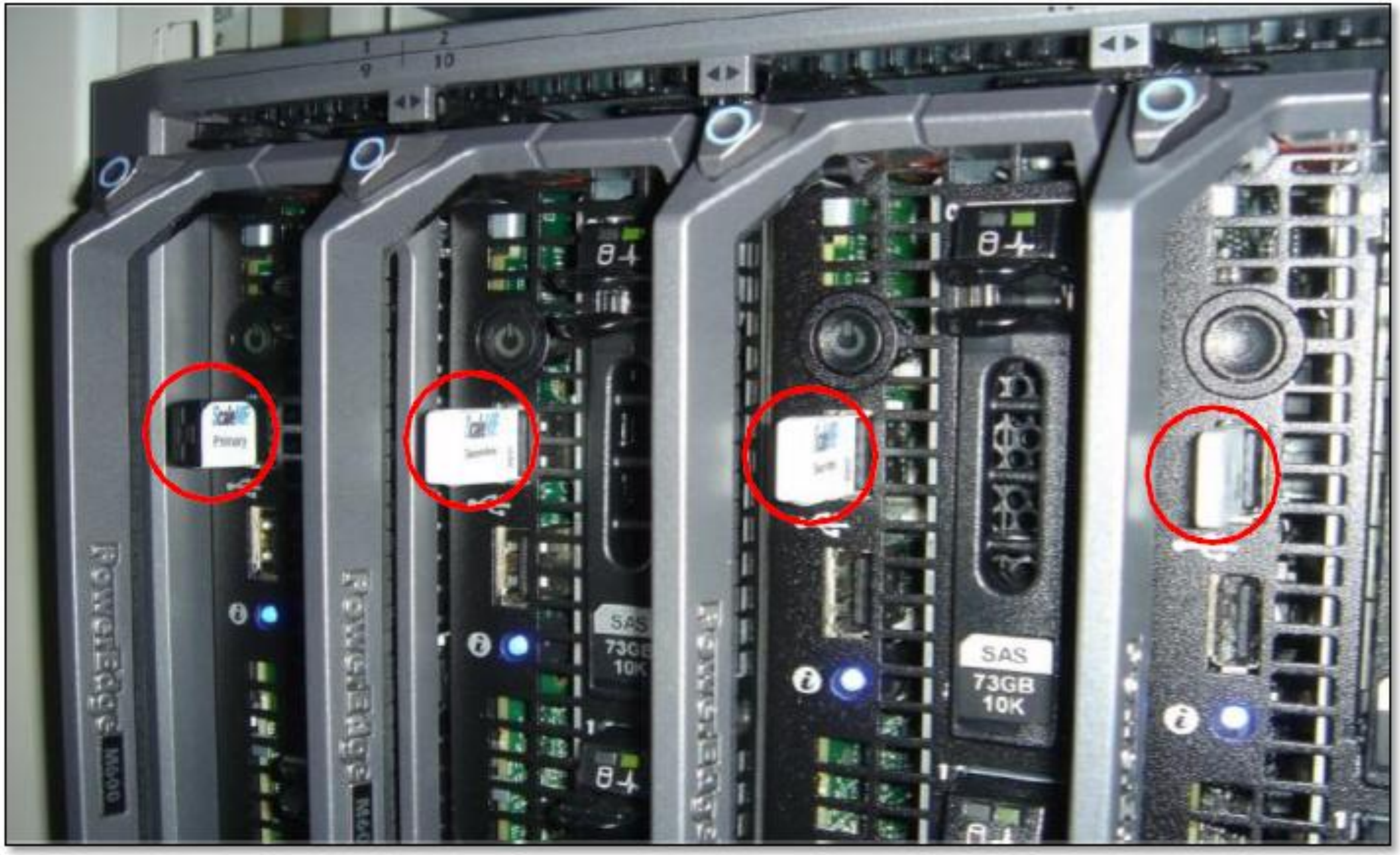
Concatenation of physical resources

(For applications requiring a superset of the physical server resources)

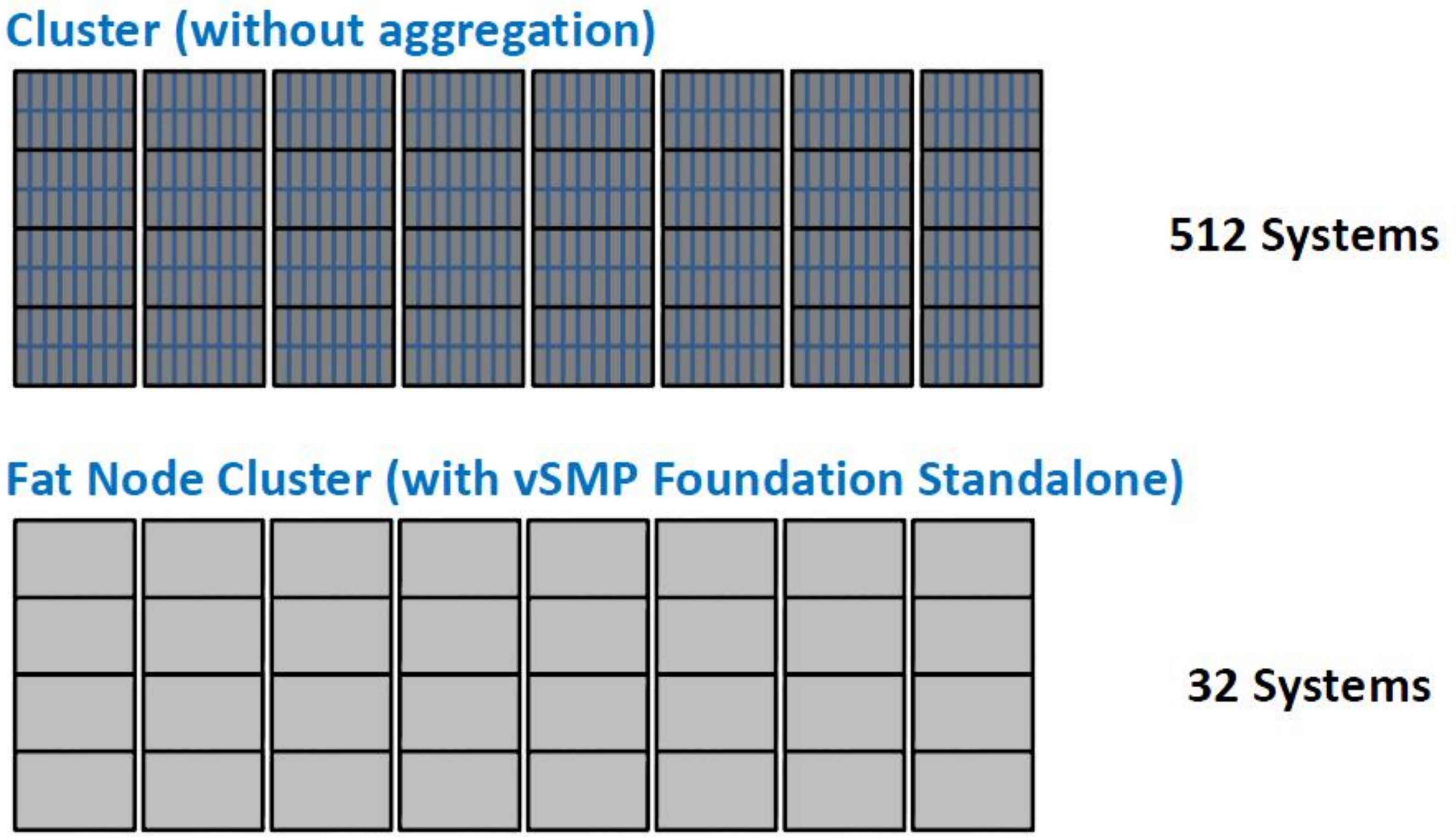


ScaleMP TM

What does it look like?



Fat node HPCC


Cluster (without aggregation)
512 Systems
Fat Node Cluster (with vSMP Foundation Standalone)
32 Systems

ScaleMP config options

- Combining 16 Dell R910 servers results in:
  - A machine with 16 TB of memory
  - And 512 CPU’s

- To save cost, smaller Dell servers can be used
  - Turn one Dell M1000e chassis into a vSMP machine
  - Up to 192 cores and 3 TB memory

- Scale up when needed, just add servers (*)

Future of HPC

- Hardware price becomes irrelevant to most usage
- Programmers will determine future
  - OpenMP, MPI, CUDA, OpenCL?
- Commercial software licenses remain expensive
- Academics have to adapt to market
- Cloud based HPC software

- We didn’t talk about storage