Honey, Who Shrinks the Supercomputer!

Jimmy (Jyh-Ming) Jong Ph.D., Tomonori Hirai, Mario Lee Ph.D.
TYAN High Performance Computing Product Group
NCHC PC-Cluster Seminar, Sept. 2006

Agenda

- (1) Computer Systems @ Crossroads
- (2) Overview of the current Commodity-Off-the-Shelf (COTS) Computer Hardware
- (3) The recent development on high performance computing system, system architecture and form factor
- (4) Other Considerations from HPC System Aspects
- (5) Who shrinks the supercomputers!
- (6) Summary and Q/A

Definition of Supercomputers

-Ranking of Supercomputers based on LINPACK benchmark GFLOPs (Billions Floating Point Operations per Second)

#1@1993 Thinking Machines CM-5, Rmax ~ 60GFLOPS

#2@1993, Rmax~30GFLOPs

2006 TYAN PSC Typhoon-1
Rmax: 40~60GFLOPS
Gbe Network, COTS 4-N Cluster

LINPACK GFLOPs is a CPU FP Intensive Benchmark.
Design Balance Hardware System (CPU, Memory, IO/Network) for most of the application

(1) Crossroads

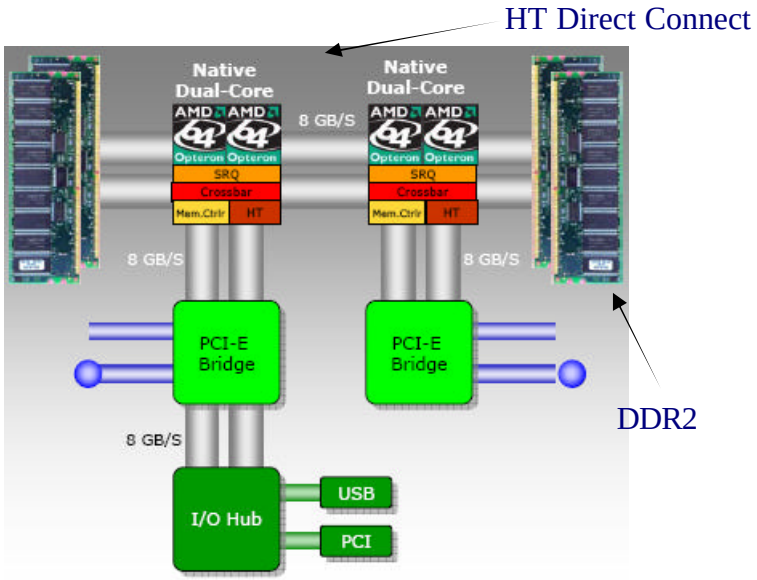
- Old Wisdom
  - Power is free, Transistor is expensive.
  - ILP via Compliers, Innovative on OOO, VLIW
  - “x” and “/” are slow, Memory Access is Fast.
  - Uniprocessor clock rate and performance: 2 x per 1.5 year
  - Performance at All Cost
- New Wisdom
  - Power Wall: GREEN. Power is expensive.
  - ILP Wall : Diminishing the investment on HW.
  - Memory Wall: 100ns cycle of Mem Access.
  - Brick Wall: CPU Cores per Chip, 2 x Cores per 1.5 year
  - TCOW and TCOP: Performance per Watt, GFlops per $1.0.

Reference: 2006 Prof. David Patterson UC. Berkeley Lectures

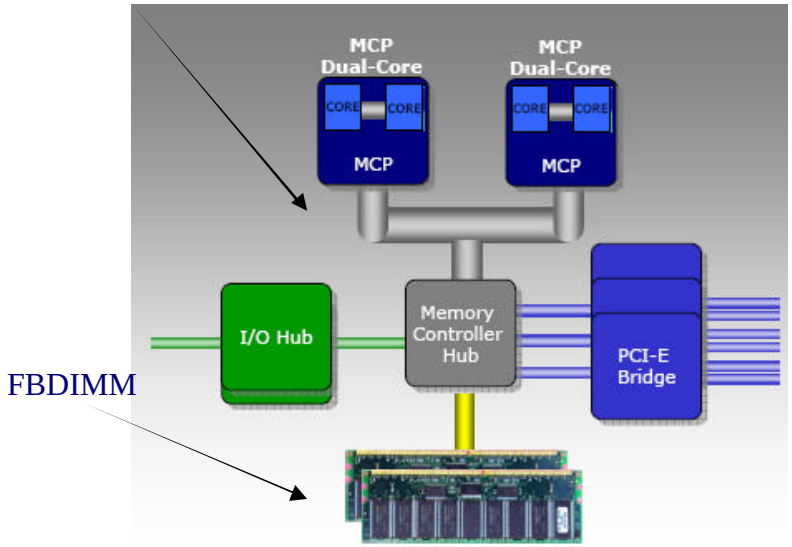
(2) COTS Hardware: CPU

- CPU: Micro-Architecture Similarity & Difference
- Performance: Micro-Architecture, System Architecture

- AMD Opteron Platform



- Intel Woodcrest Platform

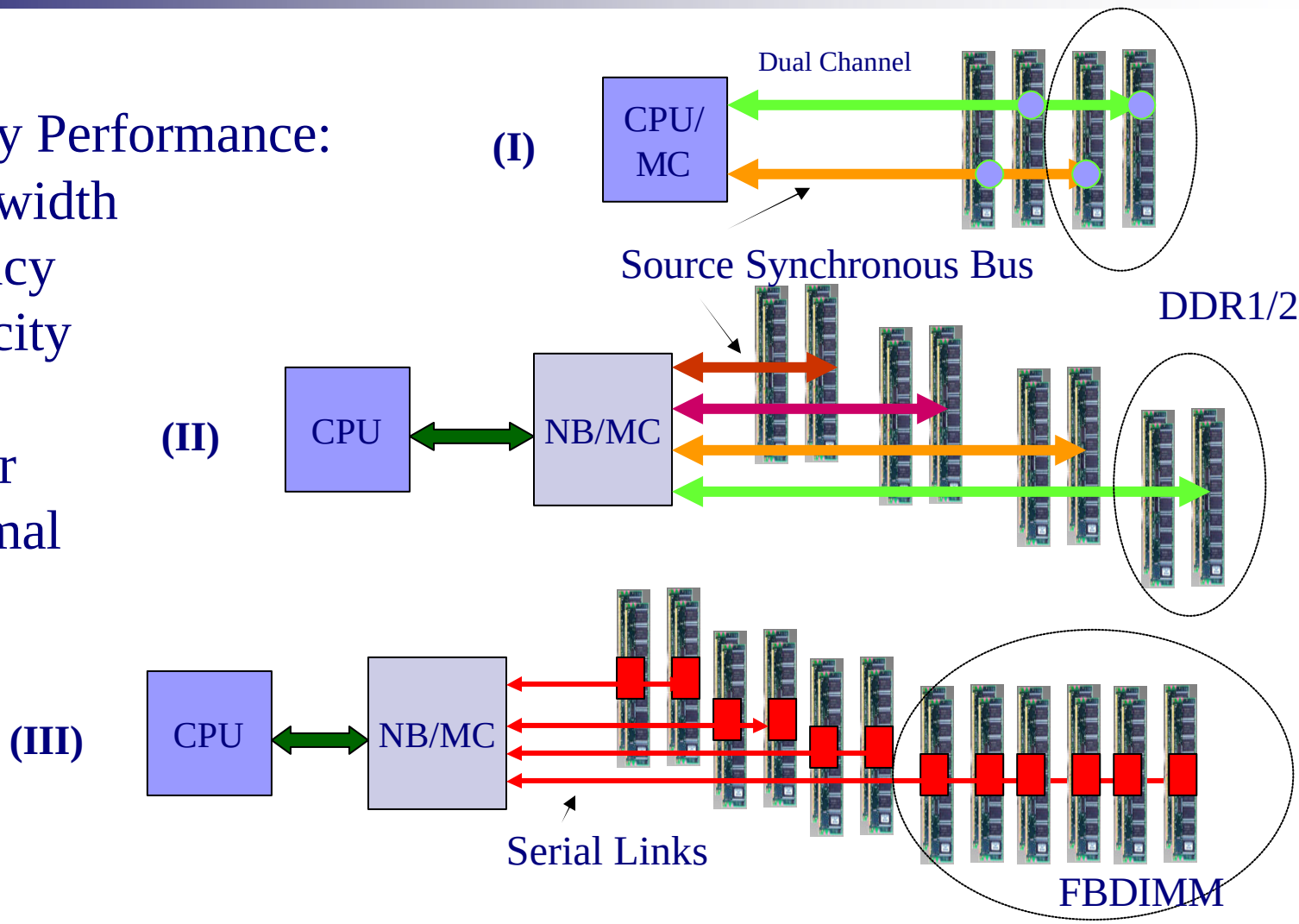


Reference: AMD and Intel materials

Memory Interface

Memory Performance:
- ✓Bandwidth
- ✓Latency
- ✓Capacity

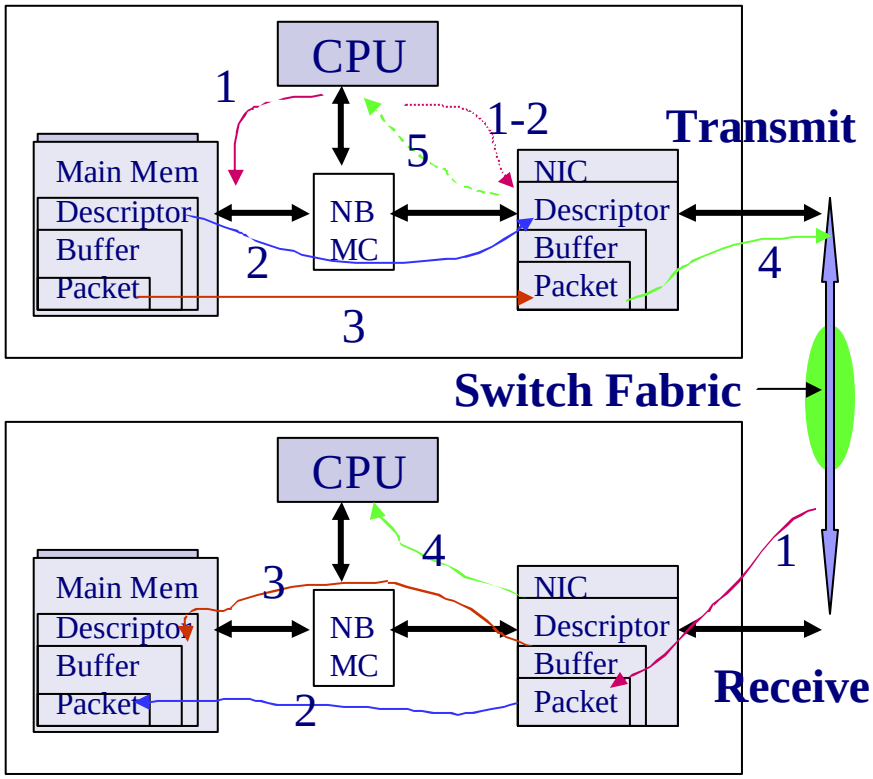
Others:
- ✓Power
- ✓Thermal
- ✓Cost



IO Interface

How the computers communicate to each others ?
And what is the impact to System Performance ?

Example: Ethernet NIC



Data movement is Expensive !

- ➢Consume CPU efforts (interrupt)
- ➢Consume FSB Bus bandwidth
- ➢Consume Memory BW/Latency
- ➢Protocol Processing is Very Expensive

IO Technology

- ✓Bandwidth
- ✓Latency

Other Improvement on NIC:

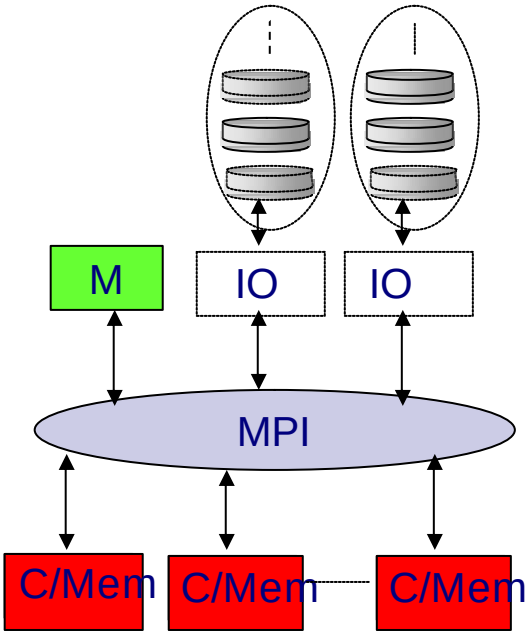
- ➢TCP/IP and Checksum Offload
- ➢Jumbo Frame
- ➢Interrupt Intervention Time

❖RDMA ?

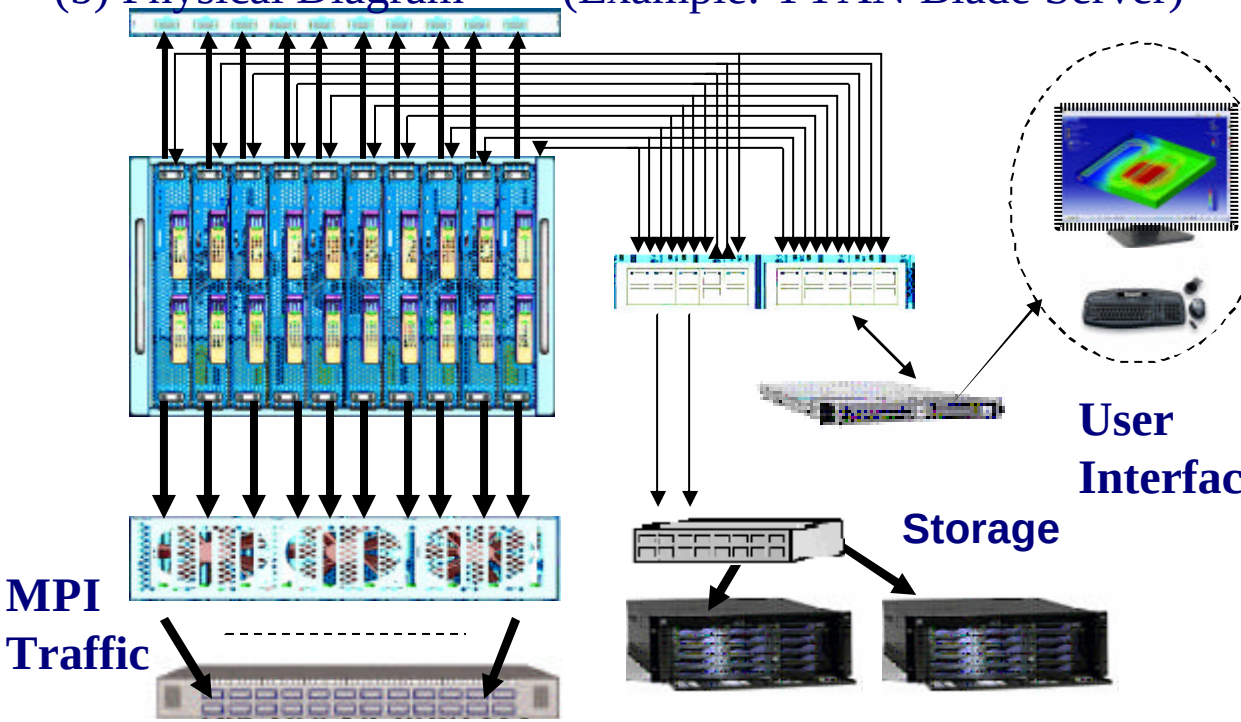
Cluster Network: Traffic Types

(1) TCP/IP for **Cluster Management**, initiating the jobs and visualizing the results
(2) MPI between CPU Nodes (MPI Interconnect Switch Fabric)
(3) Data Storage for providing data into Cluster and from the Cluster into a repository. (Either File or Block Storage via Ethernet or FC or IB)

(a) Functional Diagram



(b) Physical Diagram (Example: TYAN Blade Server)



Network IO and Switch Requirements

- Higher IO Throughput, Bandwidth
- Improve CPU Utilization
- Lower Latency
- Example: Gbe Ethernet (1000Mb/s), Infiniband (x4, 10Gb/s), 10G Ethernet

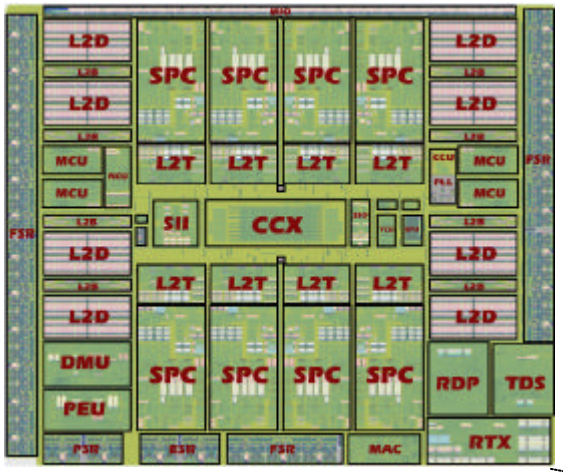
	Gbe Eth (with L2 SW)	x4 IB/PCI-x (133Mhz)	**x4 IB/PCIE (x8 PCIE)**	10G/TOE (x8 PCIE)
Throughput	**120MB/s**	740 MB/s	**880 MB/s**	0.9~1.1GB/s
Latency	**50u~70us**	5.0us	**< 5.0us**	<4.0us
CPU Utilization	**50%**	<10%	**<5%**	(~IB)

Reference: OSU, CISCO, LNLL and IBM

(3) Recent Development

- Performance per Watt: SUN CMT uSPARC SMP

SUN Niagara-2



2 x 32 single uSPARCii Systems

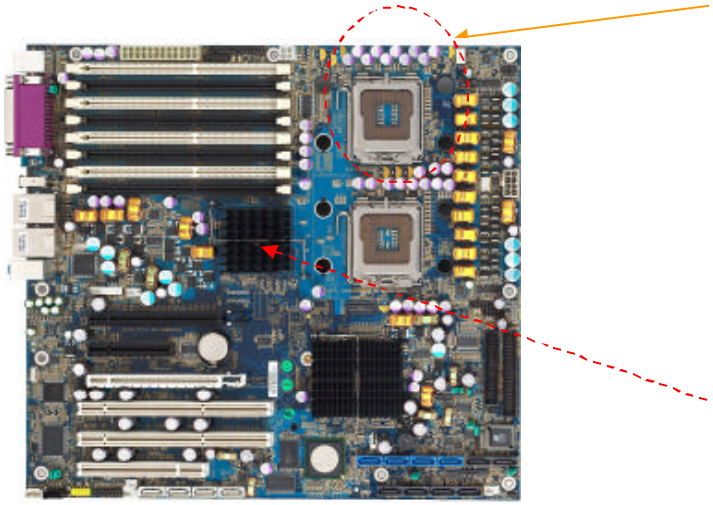
8 Cores, 8 threads per Core
Shared 4MB L2, 8 Banks
Integrated 4x FBDIMM MC – Dual Channel
Integrated 2 x XAUI/1Gbe MAC/PHY
Integrated 1 x PCIE x8 Port

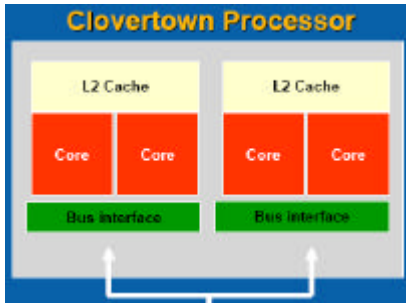
Web/App Throughput
- Not for HPC!
- Wait to see "ROCK"

Reference: Hot Chip 2006 Conf

Multi-Core Microprocessor

- Cores per Dollar: x86 Platform Intel Quad-Core, 8way SMP





Intel

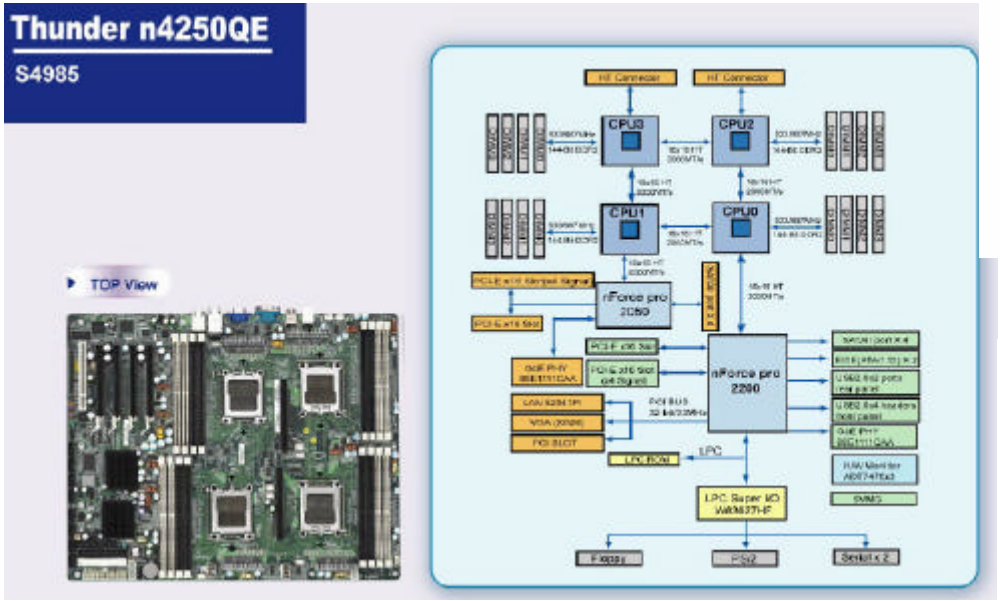
Snoop Filter to reduce
Cache Coherence
Traffic, improve 3%~5% Avg

- Dual Core Woodcrest with Dual FSB is leading SPECin_2000 and SPECin_rate_2000, competitive with AMD Socket_F/DDR2 on 2x2 Platforms.
- Socket_F 2x2 Platforms are edging on SPECfp_2000 and SPECfp_rate_2000.
- FSB Cache Traffic, Mem BW/Latency ? Large Data Set vs Cache-friendly ? HPC or Web ?

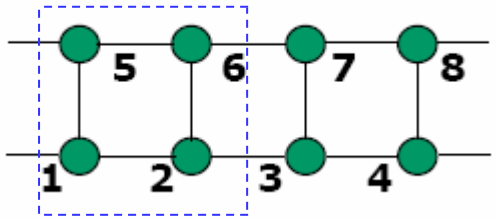
Reference: TYAN/www.tyan.com and Intel

x86 NUMA SMP System

- Commodity Mid-range SMP: AMD 4, 8 Sockets

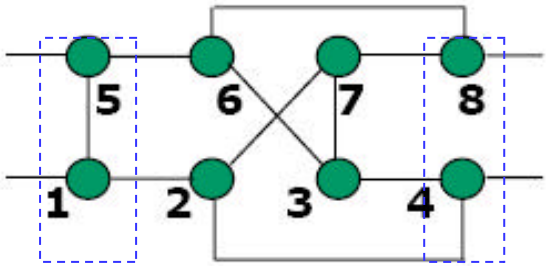


Topology (I)



2	4-hop
6	3-hop
10	2- hop
10	1-hop

(II)



0	4-hop
6	3-hop
12	2- hop
10	1-hop

Latency

•8-way/16-way (Cores) SMP System, with Dual Core CPU.

➢Non Uniform Memory Access (NUMA) – binding process of Cores with Local Memory

➢Network Bisectional Bandwidth ?

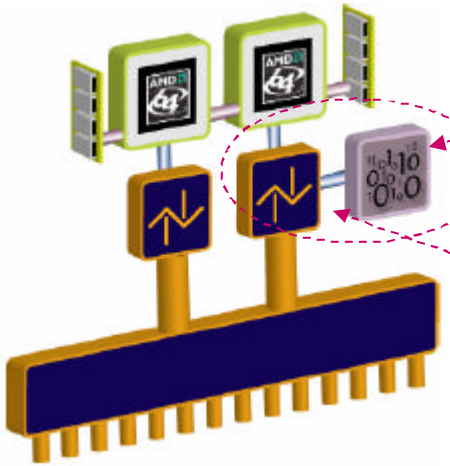
❖16Way/32Way (Cores) AMD Platform ? Network and Memory Bandwidth/Latency?

Reference: TYAN and AMD

IO, Network, Switch and Interconnect

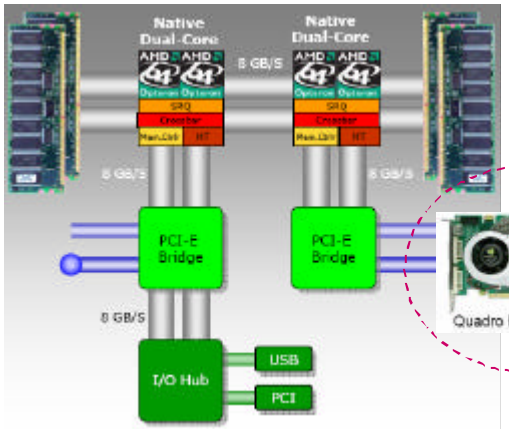
- Gbe NIC/TOE and High port count/low cost Gbe L2 Switch
- Increasing Infiniband throughout
  - Double Data Rate Infiniband: 5.0Gb/s per link, x4 HCA Card (20Gb/s raw bandwidth) and IB Switch (24 Ports * x4 channel * 2 Tx/Rx * 5.0 Gb/s peak raw BW).
  - FAT Pipe vs DDR: x8 IB, x12 IB ?
  - DDR x8 PCIE or SDR x16 PCIE to keep up with HCA Port throughput ?
    - Ramping Up ?
  - 10G NIC/TOE and 10G Switch Fabric ? Can it be < $200. per Port NIC, < $30. per Port Switch, similar to IB?
  - Optical Interconnect for Computer Room/Data Center ? Can it be ~ $5. per Gbs, similar to Copper?

Co-processor and Accelerator



Connection Bridge

- FPGA Co-processor: Execute Instruction dispatched by CPU
  - Reconfigurable Computing
  - Tightly Coupled to CPU
  - Application Specific



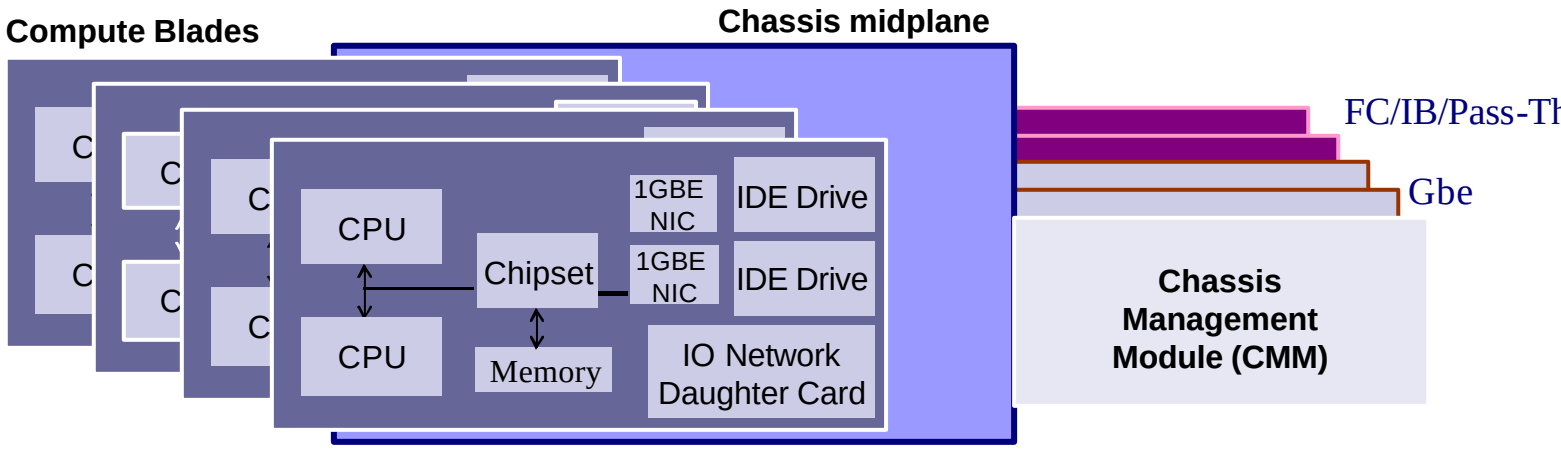
- Accelerator: Appears as a Device on the Bu
  - Can be ASIC, FPGA and Standard Device (GPGPU)
  - Controlled by Registers and synchronized with CPU to solve the problem

Reference: GPGPU.Org, Cray, IBM- Cell, ClearSpeed, XtremeData, DRC etc

Highly Integrated Server Form Factor

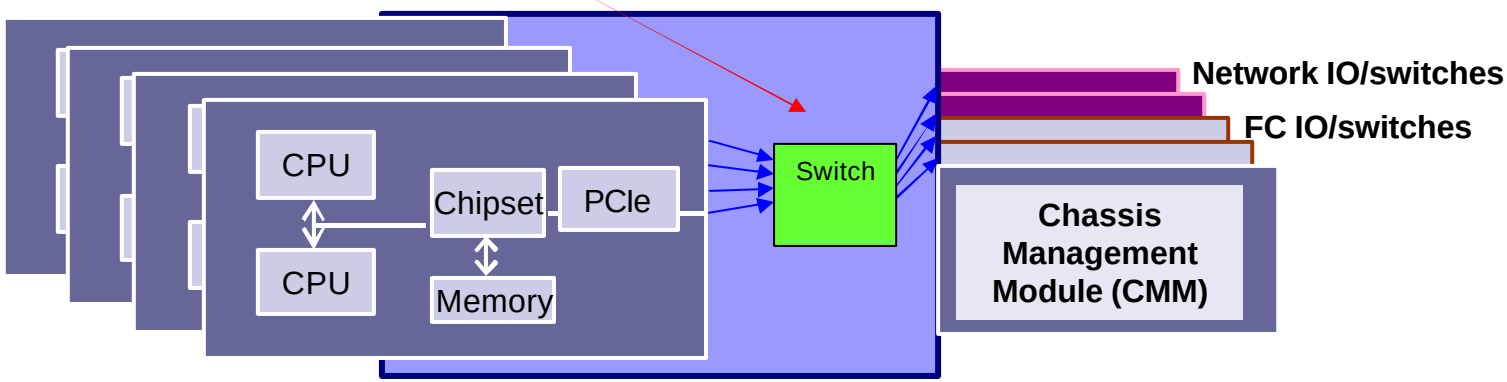
- Blade Server: Integrated Multiple Switch Fabric

Today



Active backplane- battle of Internal Switch Fabric

Future



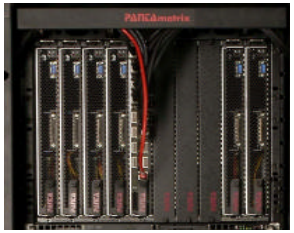
ference: MSFT and BladeSummit.org

HPC Blade Server

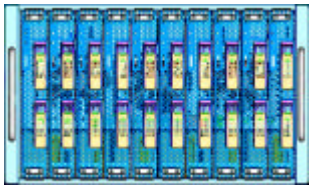
- **Commodity Blade Server vs Rack Mount Server**
  - **Benefits and Disadvantages ?**
- **Due to the limited IO bandwidth and Cost, most of commodity Blade Servers are not aimed at "HPC Market". (SGI, Cray, NEC and Hitachi have their own large SMP Blade or Brick System with custom NB chips)**
- **There are some vendors delivering commodity HPC Blade Servers.**

IBM BladeCenter H
Dual-socket, 2 x4 IB Fabric
PCIE links thru Midplane

Panta System Matrix-1
4/8-Sockets, 2 x4 IB Fabric
IB links thru Backplane



Tyan's ODM/OEM project
Dual-Sockets, Flex ext Fabr
PCIE Links thru Midplane



(4) Other Considerations for HPC

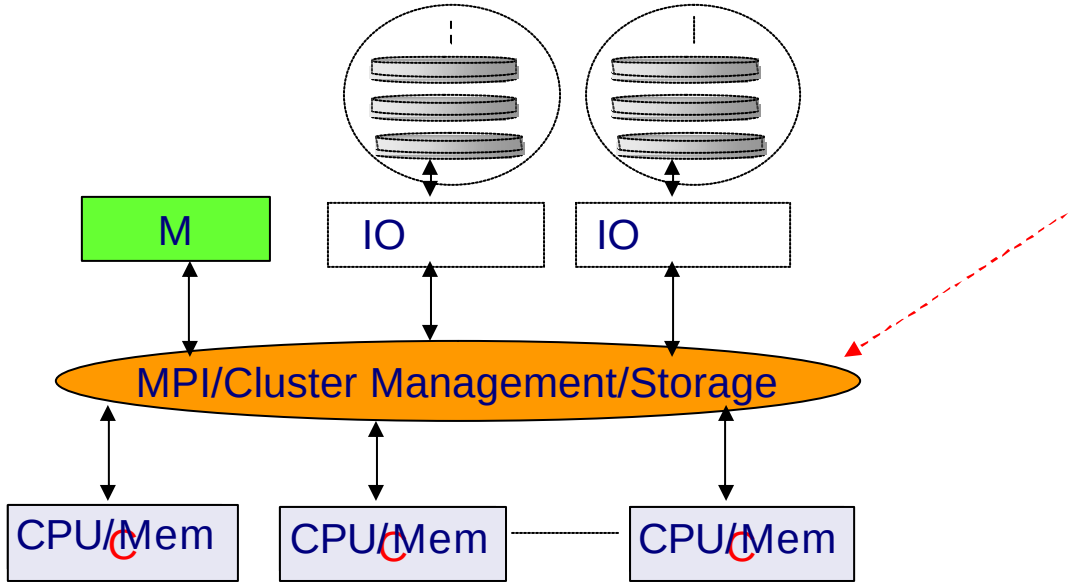
- Cooling and Thermal Challenges:
  - Inside the System: Multi-core high power CPUs (>100W) and GPU (>100W)
  - Outside the System: Data Center/Computer Room Cooling: 3 or 4 x 10kW per Rack ?
  - Cost for extra higher density in Data Center ?
- Power Delivery Challenges:
  - Inside: High efficient and high density PSU, VRM, Physical
  - Maintain commodity cost: $10c per W ?
  - Outside: Data Center Power Delivery on 40kW per Rack ?
- System Management, Diagnostic, Fault Alert and RAS requirements?

Scalability of Cluster System

- Scale Out : Clustering more systems thru Network Switch (Distributed Memory)
- Scale Up: Increasing CPU/CORE per Domain (Shared Memory)
- SMP Cluster System: 4x2x2 vs 2x4x2 ? Programming Model on intra-node and inter-node ?

- Dual-Socket
- 4-Socket
- 8-Socket
- 16-Socket

- Dual Cores
- Quad Cores
- 16 Cores



- 2 Nodes
- 4 Nodes
- 8 Nodes
- 16 Nodes

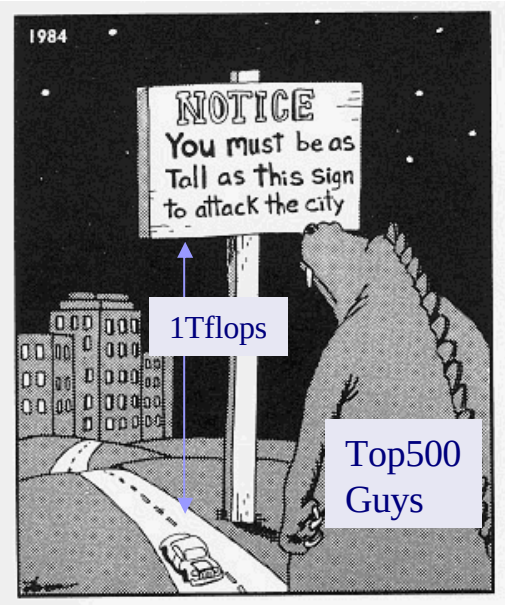
More…

Not mention in this talk

- SW/OS, Kernel Tuning for HPC ..
- Complier, Debugger…
- Cluster Middleware, Cluster deployment etc..
- API, Programming Model (MPI, Shared Memory)…
- Network Topology, SW Stack and Congestion Tuning ..
- Storage Network, SW Stack and File Systems ..
- Others

(5) Who Shrinks the Supercomputer?

- Not Aim at Top500 SC race.
- Provide COTS Cluster with the computational power of Supercomputer in a faction of the Cost ($xxx.x/Gflops).
- Without compromising the system performance (Balance System: CPU, Memory BW/Latency, IO Network).
- Speed up the SW development before sending it to Data Center SC.
- Supercomputer dedicated for a person or a group, manage your computing jobs.
- 3+ vendors including TYAN- PSC



Reference: Prof. Jack Dongarra, Univ. TN

Tyan HPG PSC Product

Gbe L2 Switch 

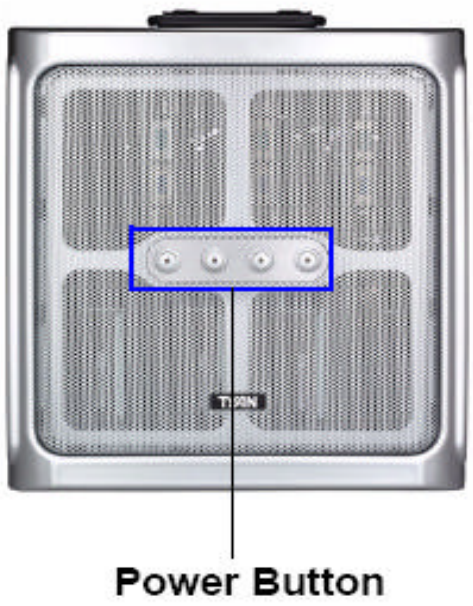
Typhoon-1
4 Nodes x Dual Socket
Ext L2 Gbe Switch
AMD/Opteron Rev.E
- 8 DIMMs per Node
Intel Woodcrest 40W/65W
- 6 DIMMs per Node

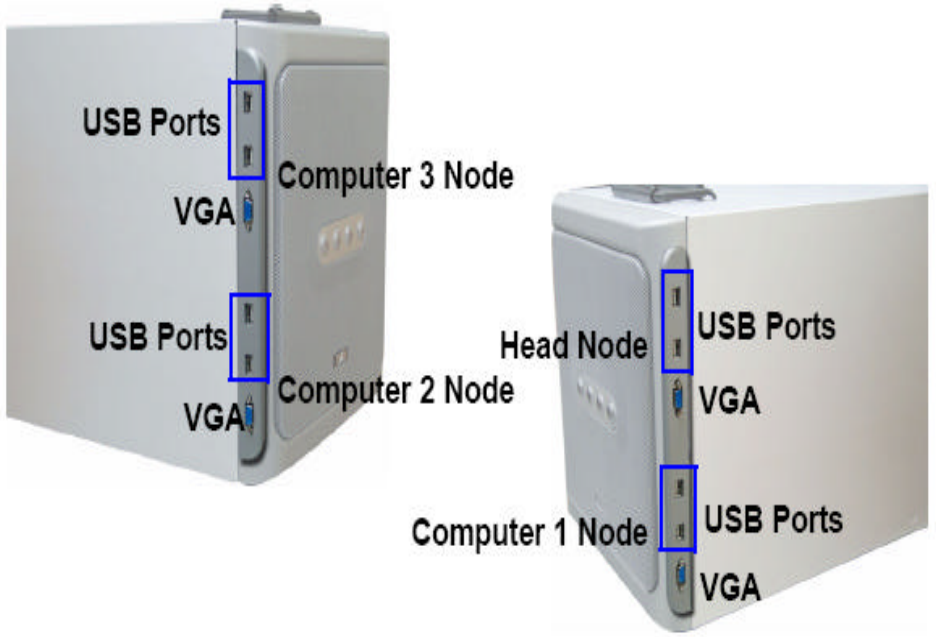
Typhoon-2
1 Head Node + 4 Compute Nodes
GPU add-in Cards (Head Node)
GPU Accelerator (TBD)
Integrated Gbe and IB Switch Fabric
Intel Woodcrest 40W/65W
Intel Covertown (TBD)
AMD Opteron Rev. E (TBD)
AMD Socket_F 65W (TBD)

✓**Offer as barebone systems, pre configured, or installed with MSFT CCS/OS.**

Typhoon-1: Overview

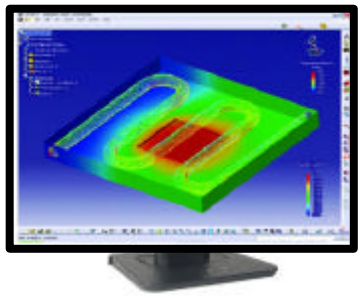
Support KM-USB/VGA per compute node for Diagnostic





Power Button per Node

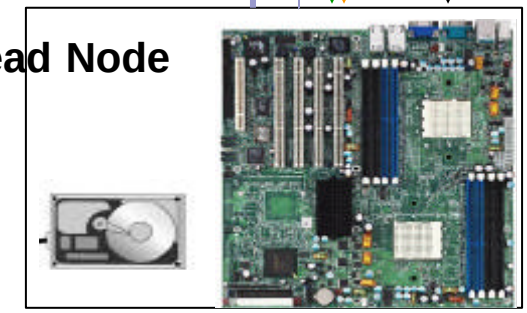
Typhoon-I Physical Block Diagram



Out side of System

100M Ethernet to Outside of PSC Network

Head Node



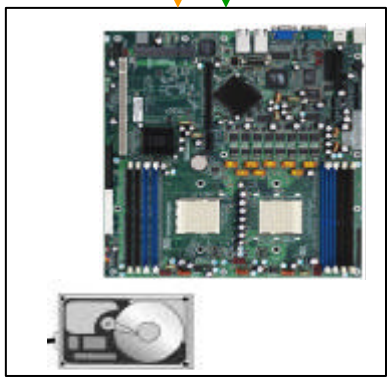
Gbe/Full Duplex

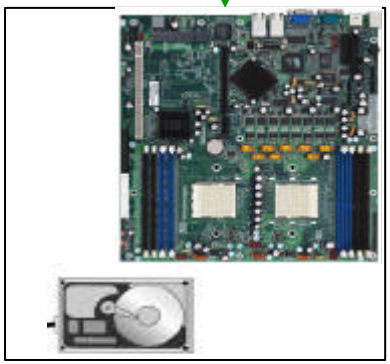
Cluster Management network

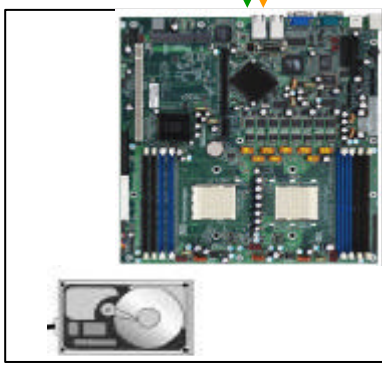
MPI

Gbe Switch

File System







Other HPG Product Line

- SMP System
- Blade Server
- Next Generation Systems (PSC, Configurable SMP and HPC Blade Server)
- Ask Tyan or check in www.tyan.com

(6) Summary

- Tremendous efforts and opportunities in COTS Cluster HPC Product development.
- Expect to see more ISV SW and Applications development supporting Cluster HPC.

END

- THANK YOU.

- Q/A ?