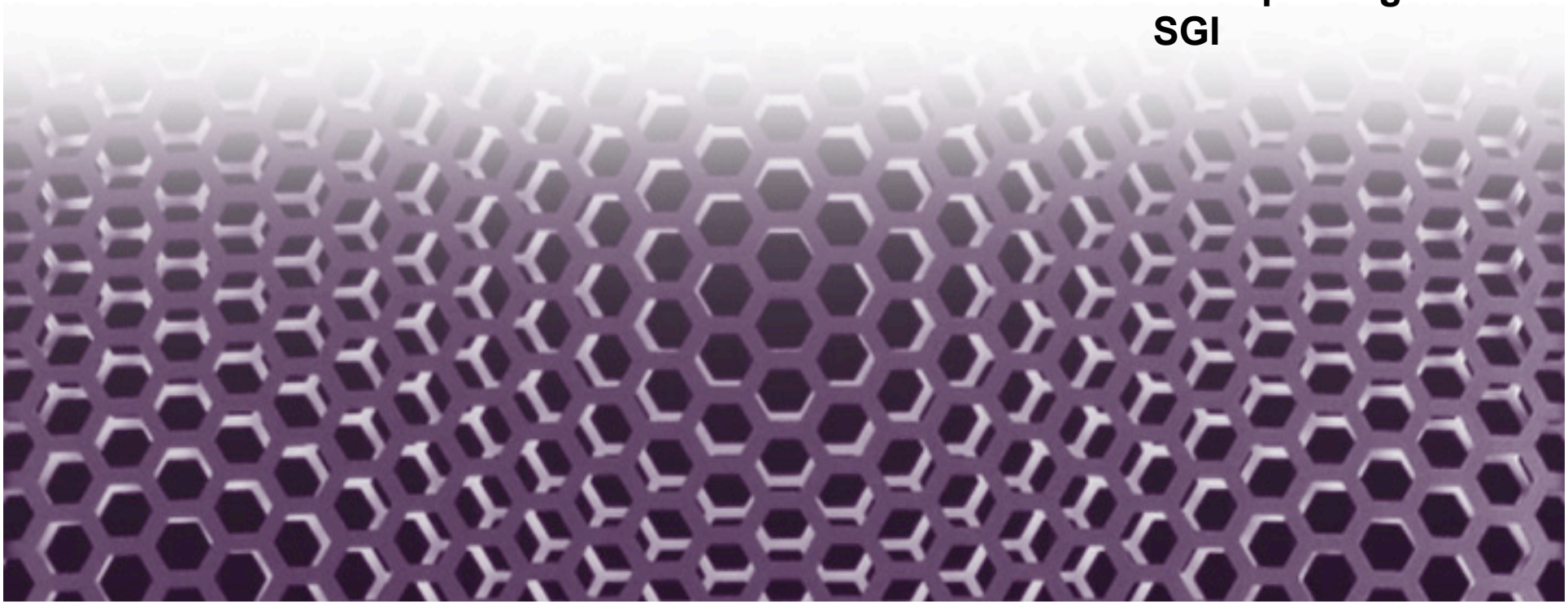
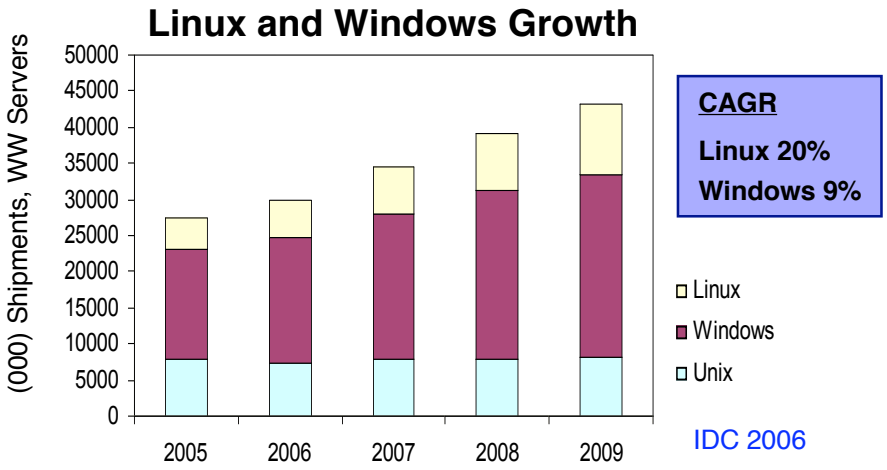
The Need For Speed

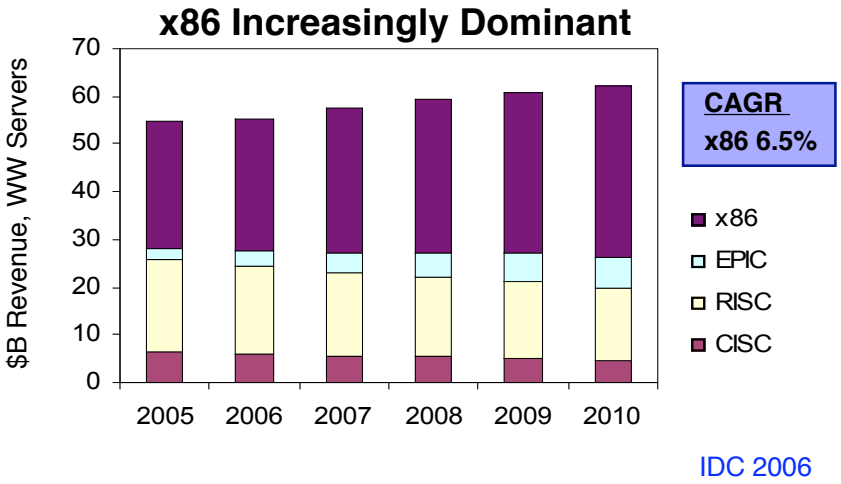
Rick Reid
Principal Engineer
SGI



Commodity Systems

- Linux
  - Red Hat
  - SUSE
  - SE-Linux
- X86-64
  - Intel Xeon
  - AMD
- Scalable Programming Model
  - MPI
- Global Data Access
  - NFS
  - NFS/RDMA
  - pNFS

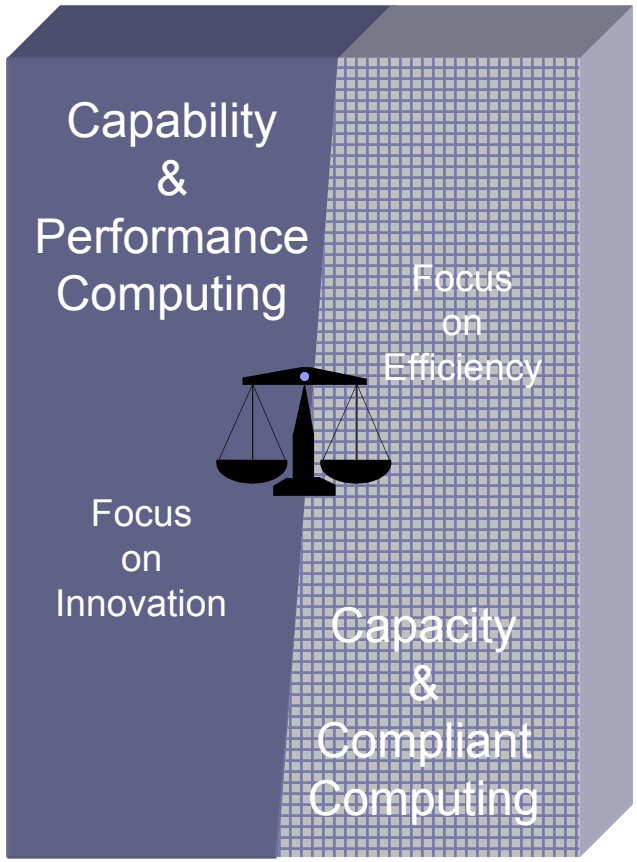




Servers: One Size Doesn't Fit All!

Workflow Characteristics

- Performance Oriented
- Data-Intensive
- Product Timelines
- Interactivity
- Rapid development cycles

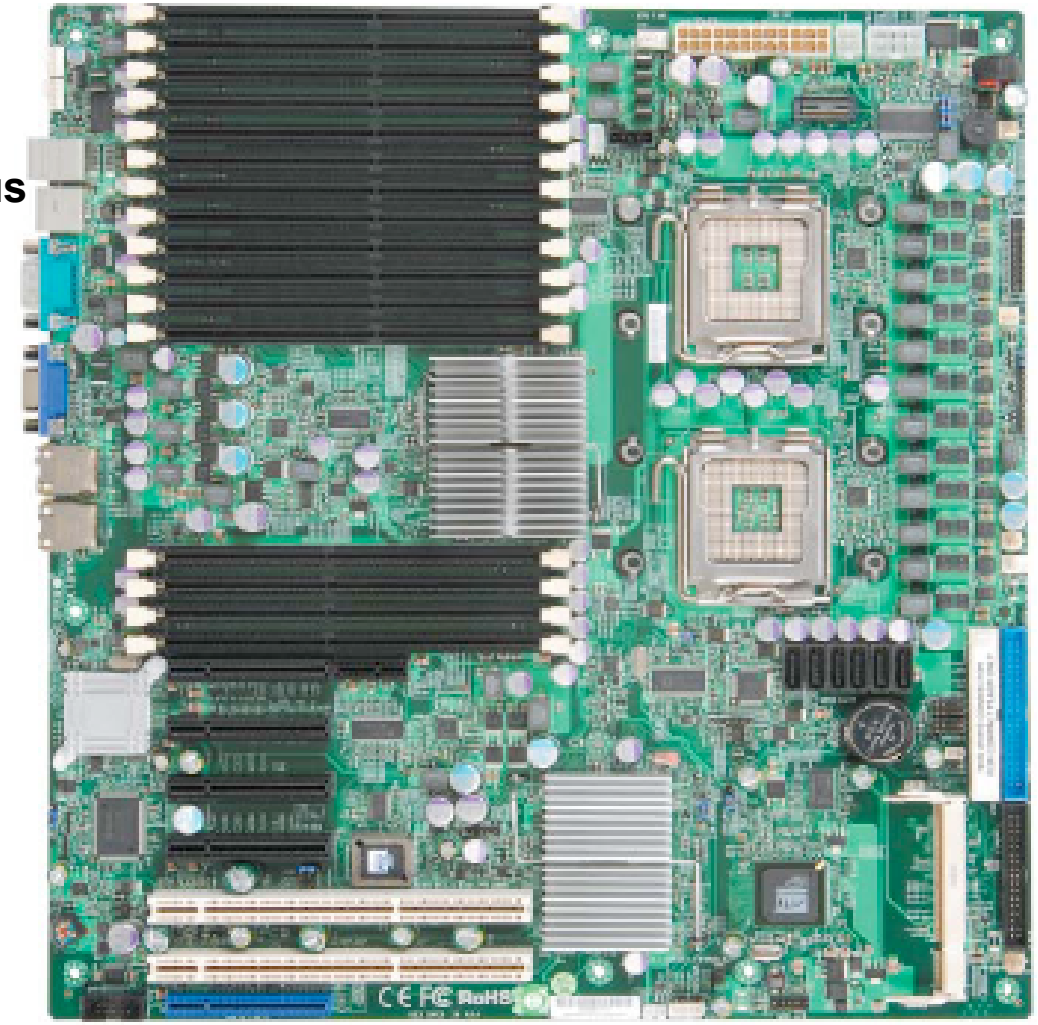


Workflow Characteristics

- Price-performance key
- Little data sharing
- Predictable Workloads
- Non-interactive
- Mature Apps

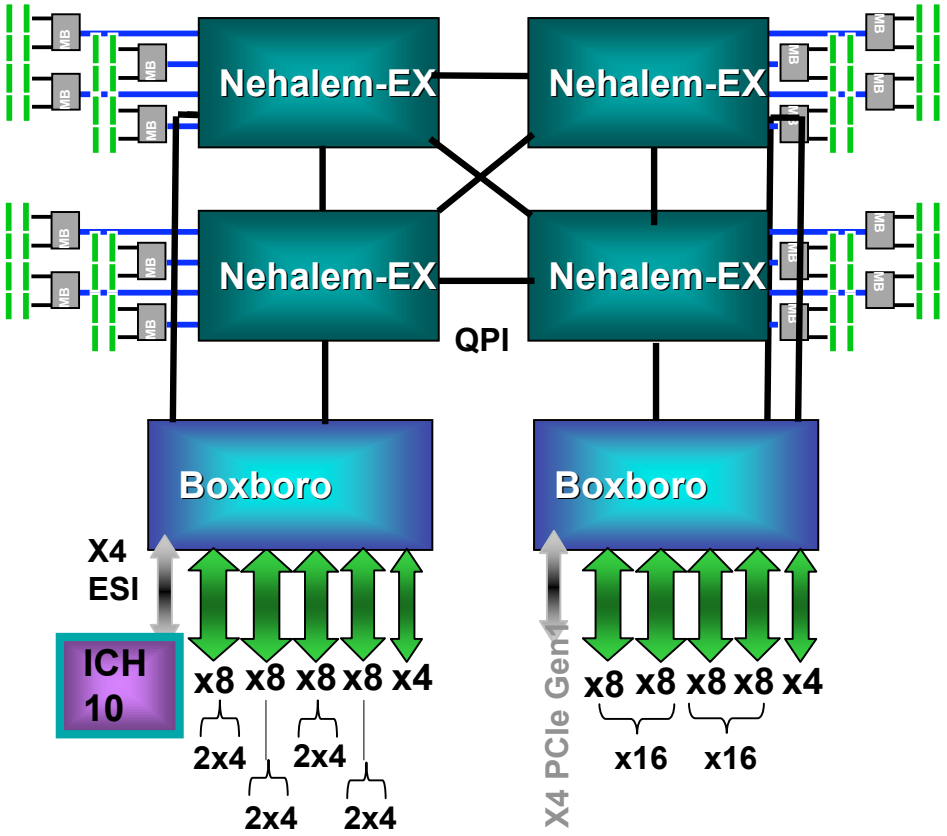
Commodity Xeon Motherboard – Today

- Seaburg chipset
  - Wolfdale and Harpertown
  - 1600 MHz or 1333 MHz front side bus
- 16 FB DIMM slots
  - 64 GB ECC DDR2 FBD
- 8 SATA2 or SAS drives
  - H/W RAID 0, 1, 5, 6, 10, 50, 60
- ATI 32 MB DDR2 Graphics
- Dual GbE interfaces
- Redundant power supplies
- 5 external I/O slots
  - 2 x PCIe x8 gen 2
  - 1 x PCIe x8 gen 1
  - 1 or 2 x PCIe x4 gen 1
  - 1 or 2 x PCI-x 133/100



Commodity Motherboard - Next Year

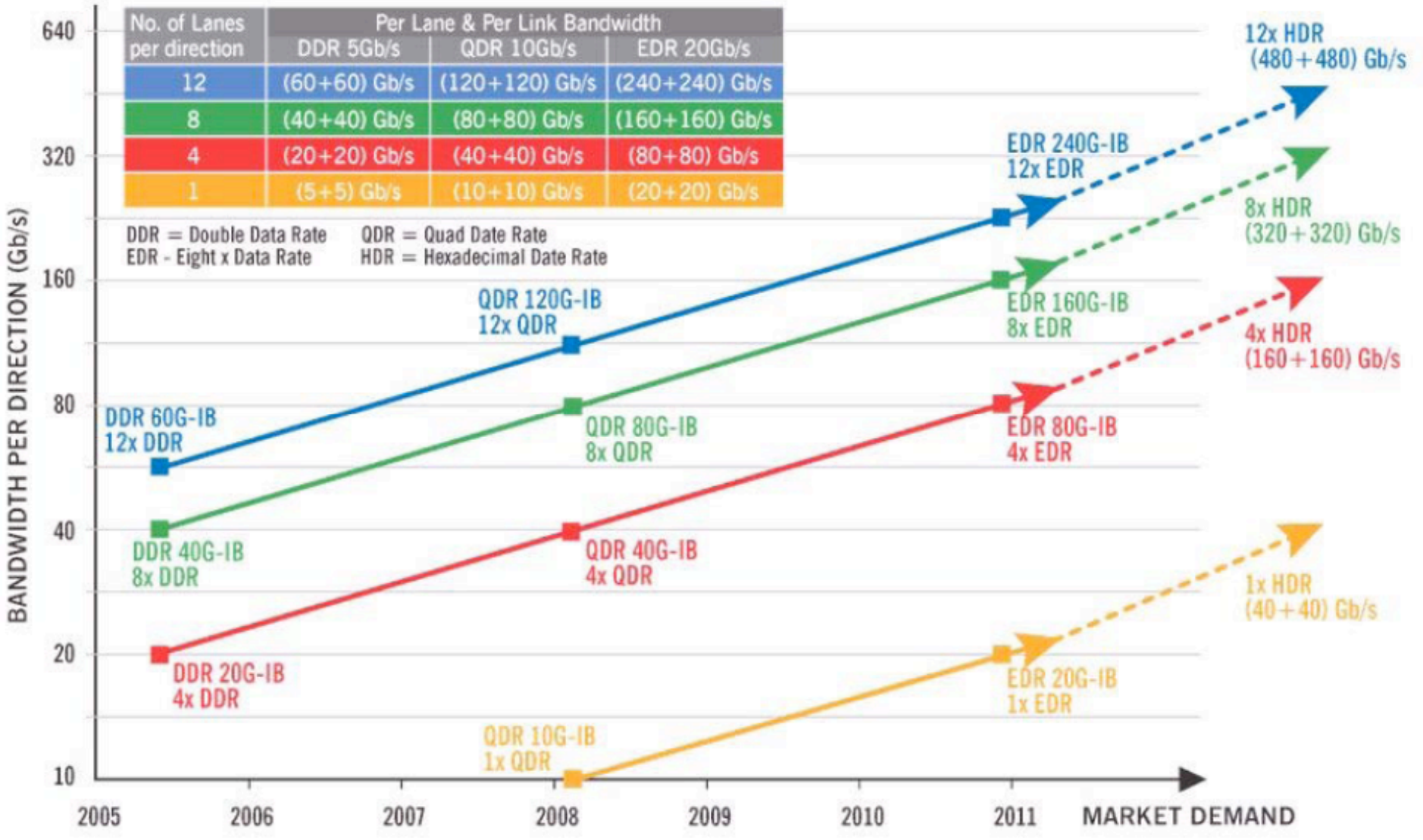
- 32 Cores
- 320 GF
- 64 DIMMs – 256 GB
- 1U



Fabrication Technology Drives the Need

2003	2005	2007	2009	2011	2013	2015	2017
90 nm	65 nm	45 nm	32 nm	22 nm	16 nm	11 nm	8 nm
1 - 2 cores	2 - 4 cores	4 - 8 cores	6 - 12 cores	8 - 16 cores	12 - 18 cores	20 - 28 cores	24 - 32 cores
3 GHz , 4 FLOPS / clock							
12 - 24 GF/s	24 - 48 GF/s	48 - 96 GF/s	72 - 144 GF/s	96 - 192 GF/s	144 - 216 GF/s	240 - 336 GF/s	288 - 384 GF/s
2 sockets per motherboard / blade							
24 - 48 GF/s	48 - 96 GF/s	96 - 192 GF/s	144 - 288 GF/s	192 - 384 GF/s	288 - 432 GF/s	480 - 672 GF/s	576 - 768 GF/s
If operations per word of input = 500 than GB/s of I/O required to prevent stalls (20% O/S overhead)							
0.31 - 0.61 GB/s	0.61 – 1.23 GB/s	1.23 – 2.46 GB/s	1.84 – 3.69 GB/s	2.46 – 4.92 GB/s	3.69 – 5.53 GB/s	6.14 – 8.6 GB/s	7.37 – 9.83 GB/s
GbE / FC4 / IB 4x DDR			10 GbE / FC8 / IB 4x QDR			IB 12x QDR, EDR, HDR	
Production			Development			Research	

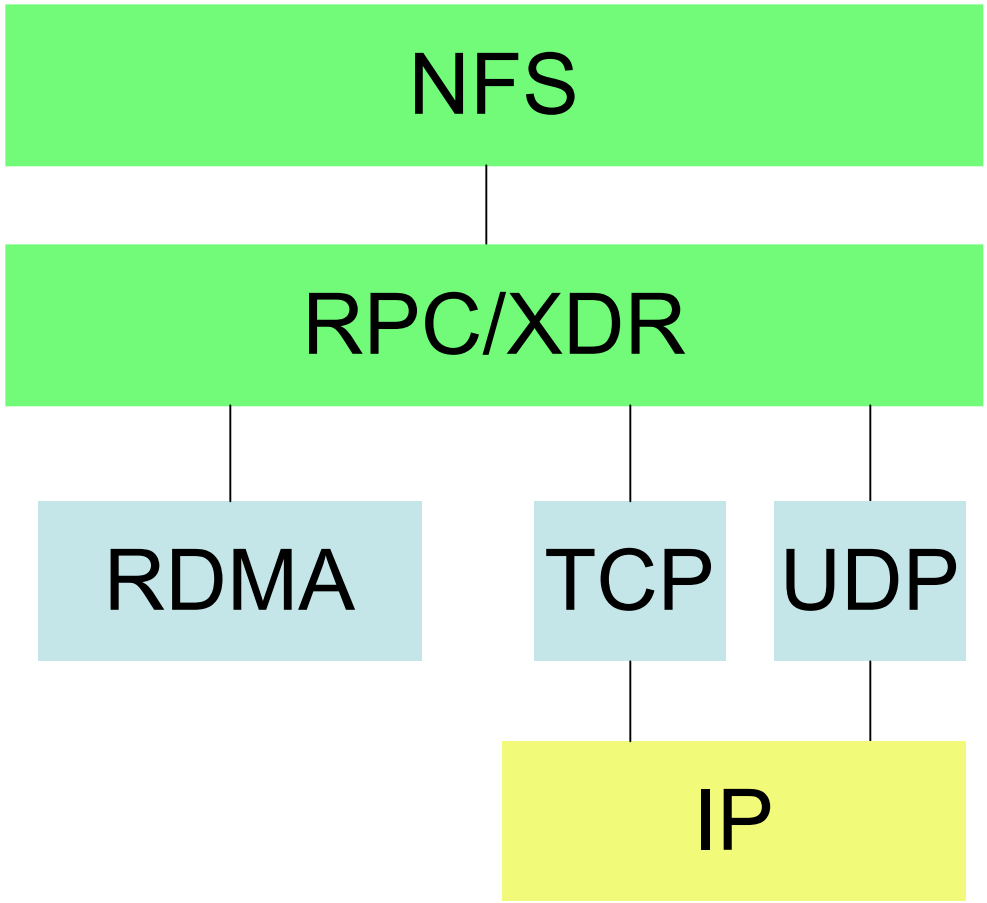
Infiniband Roadmap



NFS/RDMA - A new NFS standard

- NFS/RDMA is provided by transport switch in the RPC layer which allows data to be transported via TCP/IP, UDP/IP or RDMA over Infiniband.
- The transport switch allows NFS to use the efficient, high bandwidth native Infiniband protocol.
- NFS/RDMA can be simultaneously intermixed with NFS over IP-based protocols on the same server and even the same files
- NFS/RDMA transfers the data from the Severs memory to the client memory without host CPU intervention
  - This allows the NFS server to serve more clients
  - This improves NFS bandwidth
  - This allows the NFS client to dedicate more CPU resources to the application
- Latest OpenFabric drivers provide most of the capabilities required
  - Current version: OpenFabrics Enterprise Distribution (OFED) 1.3

What is NFS over RDMA?



NFS/RDMA

- IB or 10 GbE connectivity between NFS server and NFS clients
- IB, SAS or FC backend to storage
- Performance
  - 10 RAIDs, 480 15K SAS drives, 96 GB RAM, 2 IB interfaces
  - NFS/RDMA IB Read: 3.5 GB/s
  - NFS/RDMA IB Write: 2.25 GB/s
  - NFS 10 GbE Read: 800 MB/s
  - NFS 10 GbE Write: 400 MB/s
- Higher aggregate bandwidth
- Lower CPU overhead - ~10% CPU load/IB interface

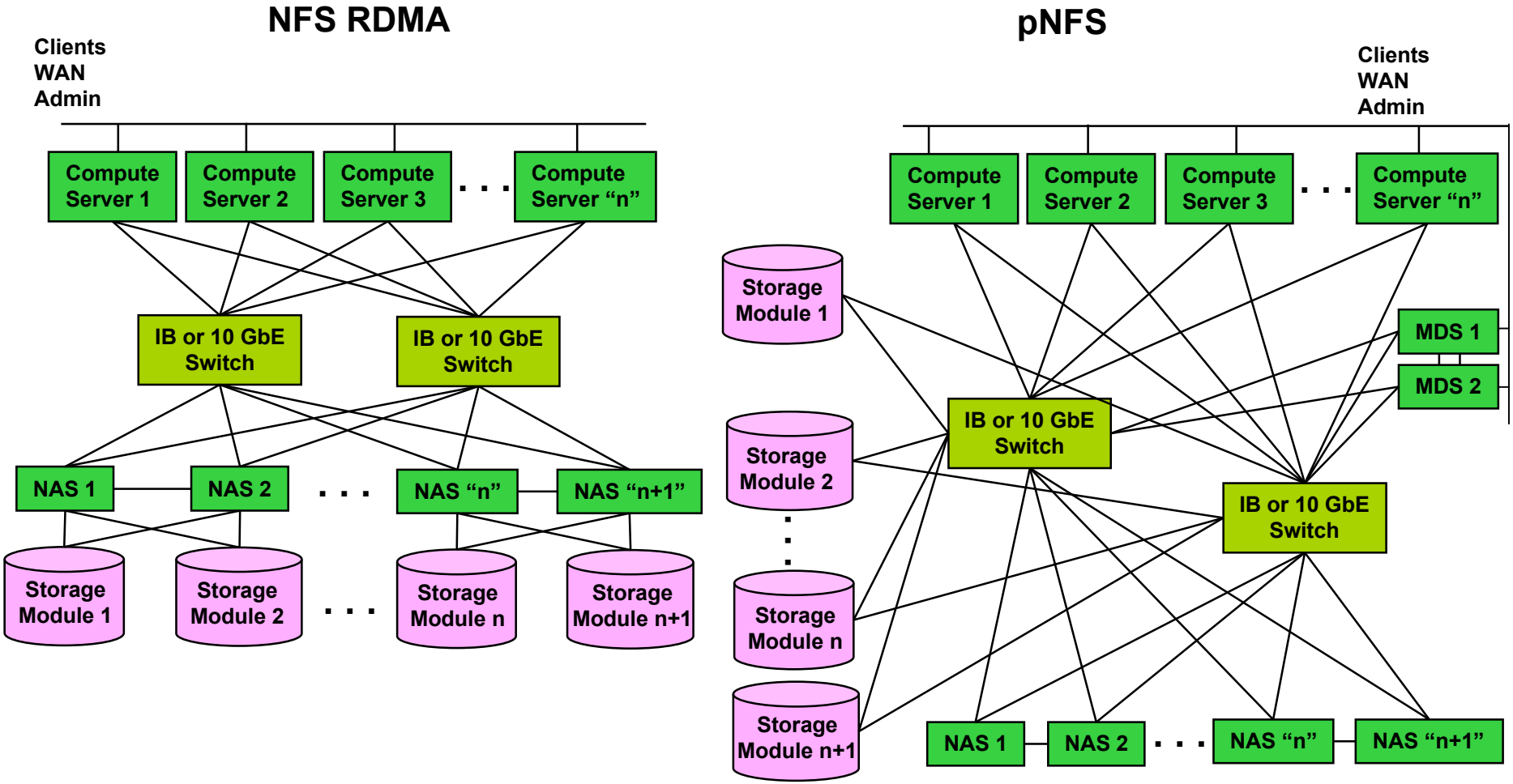
pNFS - Future Technology

- Parallel NFS: Extension of NFS 4.x Standard
  - Expected draft release in NFS 4.1
- Dr. Garth Gibson, professor at CMU, wrote the Internet Draft:

  "pnfs-problem-statement"

  www.pdl.cmu.edu/pNFS/archive/gibson-pnfs-problem-statement.html
- File base sharing
- Uses metadata servers
  - NFS semantics for metadata access
  - Metadata server points client to the storage node
  - Supports Block, File System, or Object Storage backend
  - Expected to be able to serve 100K+ of clients
- Multiple storage and system suppliers have stated their plans to support pNFS

NAS Storage


NFS RDMA
Clients
WAN
Admin
Compute Server 1
Compute Server 2
Compute Server 3
Compute Server "n"
IB or 10 GbE Switch
IB or 10 GbE Switch
NAS 1
NAS 2
NAS "n"
NAS "n+1"
Storage Module 1
Storage Module 2
Storage Module n
Storage Module n+1
pNFS
Clients
WAN
Admin
Compute Server 1
Compute Server 2
Compute Server 3
Compute Server "n"
Storage Module 1
Storage Module 2
Storage Module n
Storage Module n+1
MDS 1
MDS 2
IB or 10 GbE Switch
IB or 10 GbE Switch
NAS 1
NAS 2
NAS "n"
NAS "n+1"

Questions

Thank You