Stovepipes to Clouds

Rick Reid

Principal Engineer

SGI Federal

Agenda

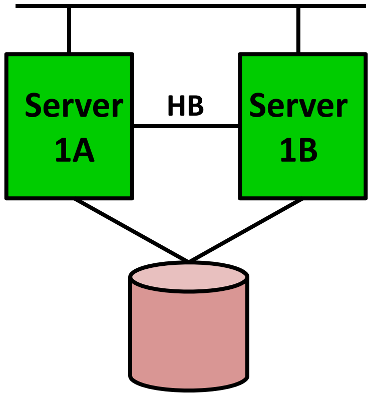
- **Stovepipe Characteristics**
- **Why we Built Stovepipes**
- **Cluster Characteristics**
- **Why we are Moving to Clusters**
- **Cloud Characteristics**
- **Why we May Not Move to Clouds**
- **Summary**

Stovepipe Characteristics

- **Numerous Servers**
  - Performance, Normal and Custom Variations
- **Primarily Global Data Access**
  - SAN or NAS
- **Proprietary File Systems**
- **Proprietary O/S**
- **Custom SMP code**
- **Expensive**

Why we built stovepipes

- **Requirements**
  - **Performance, performance, performance**
    - Fastest servers available
      - 32 to 64 sockets / cores, < 50 GF
      - 1 GB memory per core
    - I/O – FDDI, HIPPI, FC, GbE, ATM
      - All < 1 Gb/s
    - Custom code (SMP)
    - Proprietary OS and file systems
    - Custom H/W
  - **Reliability**
    - Dual capture
    - 2n server redundancy
    - Single function, hot standby



Cluster Characteristics

- **Numerous Servers**
  - Performance and General node types
  - Improved node packaging
  - Fewer, smaller, faster Servers
  - n+m redundancy
- **Primarily Global Data Access**
  - SAN or NAS
  - Faster networks
- **Proprietary or Open Source File Systems**
- **Linux**
- **Custom SMP code (for ground stations anyway)**
- **Less Expensive**
  - Power, cooling, floor space, maintenance

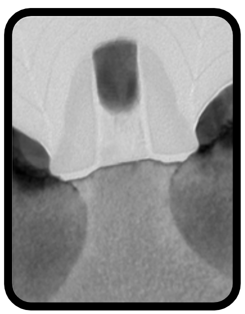
Why we are Moving to Clusters

- **Numerous Servers**
  - Improved node packaging
  - n+m redundancy
- **Faster Networks**
  - 10 GbE moving to 100 GbE
  - IB FDR moving to EDR
- **Lustre / NFS**
- **Linux**
- **Existing SMP code ports easily**
- **Less Expensive**
  - Power, cooling, floor space, maintenance

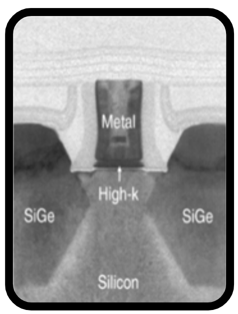
Intel's Tick/Tock Roadmap

Tick – Lead vehicle on new manufacturing process, modest change
Tock – Opportunity for significant change

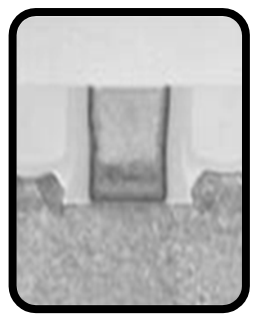
Merom	Penryn	Nehalem	Westmere	Sandy Bridge	Ivy Bridge	Haswell
NEW Microarchitecture 2006	Compaction/ Derivative 2007	NEW Microarchitecture 2008	Compaction/ Derivative 2010	NEW Microarchitecture 2012	Compaction/ Derivative 2013	NEW Microarchitecture
65nm Tock	45nm Tick	45nm Tock	32nm Tick	32nm Tock	22nm Tick	22nm Tock



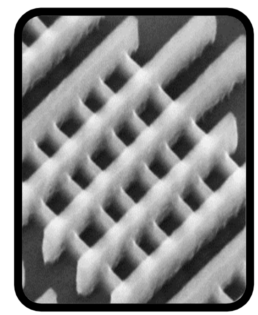
2nd Gen.
SiGe
Strained Silicon



Invented
Gate-Last
High-k
Metal Gate



2nd Gen.
Gate-Last
High-k
Metal Gate



First to
Implement
Tri-Gate

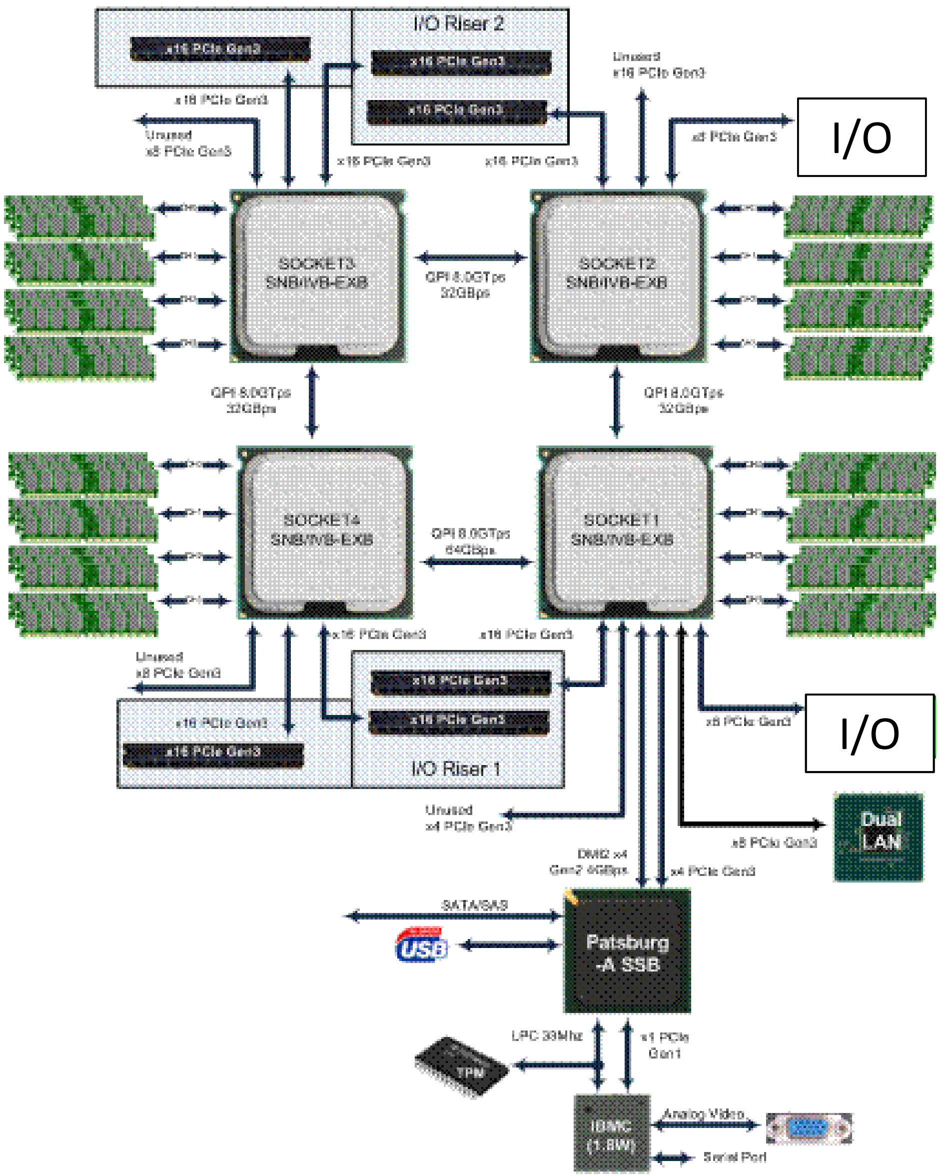
Intel EP Socket Roadmap

2010	2012	2013	2014
Tick	Tock	Tick	Tock
Westmere-EP	Sandy Bridge-EP	Ivy Bridge-EP	Haswell-EP
Shrink		Shrink	
Same uArch	New uArch	Same uArch	New uArch
Cores/die ↑	Cores/die ↑	Cores/die ↑	Cores/die ↑
Core freq →	Core freq →	Core freq →	Core freq →
DIMM size ↑	DIMM size ↑	DIMM size →	DIMM size ↑
Mem speed ↑	Mem speed ↑	Mem speed ↑	Mem speed ↑
	PCIe gen 3		
	8 FLOPs/Clock		FLOPS/Clock

Cluster - General Server

- 4 servers in a 2U node (hot-pluggable)
- Intel Xeon SB 2S-EP
  - 16 cores / 32 threads per node
  - 371.2 GF per node
  - 1.485 TF per server
- 512 GB memory capacity per node
  - 16 DIMMs per node (1600MHz)
  - 2TB memory capacity per server
- PCIe Gen3 I/O
  - One x16 (low-profile) slot per server
  - 4 2.5” drives per server
- Redundant power supplies

Cluster - Performance Server



- Intel Xeon Sandy Bridge 4S-EP
  - 32 cores / 64 threads
  - 742 GF
  - Up to 130W support
- 1.536 TB memory capacity
  - 48 DIMMs DDR3 (1600MHz)
- PCIe Gen3 I/O
  - Two PCIe3 x48 Risers
    - Four x16 slots (FLFH or HLFH)
    - Two x16 internal slots (HLFH)
  - Two I/O x8 modules

Cloud Properties - Key Requirements

<u>Lowest platform cost must also achieve these goals:</u>

1. Unlimited scaling without interruption
   - The Cloud must be expandable seamlessly

2. No down-time = 100% Availability of service
   - No one will TRUST a cloud if it goes down

3. Zero lost data
   - No one will TRUST a cloud if it loses data

4. Cost-of-service must be an order of magnitude less than the traditional compute-data approach.

5. Security must be acceptable for the users information
   - Trust is mandatory therefore – security tools must provide higher security than a closed system has ever had to deal with.

Internet Cloud Characteristics

- **No RAID Cards, all storage is JBOD**
- **No virtualization**
- **Cloud providers are driven to the lowest cost of ownership**
  - Power cost
  - Footprint cost
  - Cooling cost
  - Purchase cost
  - Labor cost for maintenance
  - Cost of upgrading hardware (all of the above) every 3 years or whenever the cost of operations exceeds the cost of upgrading/performance
- **Balanced hardware configurations: cores to spindles to GB Memory**
  - Keep a general purpose consistent hardware infrastructure across all data centers
  - There should be no difference in performance and jobs can be reliably moved to any server
  - Scalability is the key to maintaining the lowest cost of ownership
- **All Remote Bootable**
- **No DVD drives in any server**
- **No extra gear of any kind in any server**
- **Memory is typically 4GB per core using 8GB DIMMS to keep power as low as possible.**
- **Cloud providers are weighing the cost , the performance and the cost of operation against the full cost of ownership over multiple years.**

Internet Cloud – Required a New Approach

Unlimited, seamless-scaling required a change

Traditional IT "enterprise" approach to compute-storage platforms:

- ***Send the data to the question for processing***
  - Pull the data into compute then return the answers to storage when finished
  - Expensive, large-redundancy-rich compute platforms run queries and processing
  - The associated storage platforms are very robust and redundant
  - Compute is compute, storage is storage and processing is done at the compute side with the data moving across a fast and redundant storage network.

Internet-Cloud platforms required a different approach:

- ***Send the question to the data, not the data to the question!***
  - Enter Hadoop/MapReduce – and all the attendant tools
  - In order to scale seamlessly the cloud required a continuing expansion of the compute and storage with standard building blocks at the lowest total cost
  - The building blocks must be added to a running system providing both compute and storage increments in a predictable and useable manner
    - All building blocks must run the same file system and OS platform
    - All storage must have maximum speed per $ spent and 100% reliability
      - Speed measured is from the CPU to the data (no networking makes that faster) – DAS
      - No RAID at the hardware level – slows down data flow to CPU
      - Software RAID at the File System level across multiple server-DAS at multiple locations

Why We May Not Move to Clouds

- **Servers with special needs**
  - Custom I/O
  - Performance nodes
  - Reliability and failover capability
- **No global I/O accessibility**
- **Interconnect generally 1 GbE or 10 GbE**

Summary

- **Technology will provide powerful enough nodes**
- **SMP code probably does not have to be ported**
- **Linux rules**
- **Open source globally addressable storage (SAN or NAS) is usually not available in a cloud**
- **Moving from a stovepipe to a cluster (FLOP for FLOP) will result in facility and maintenance savings over 3 years that will pay for the replacement systems**

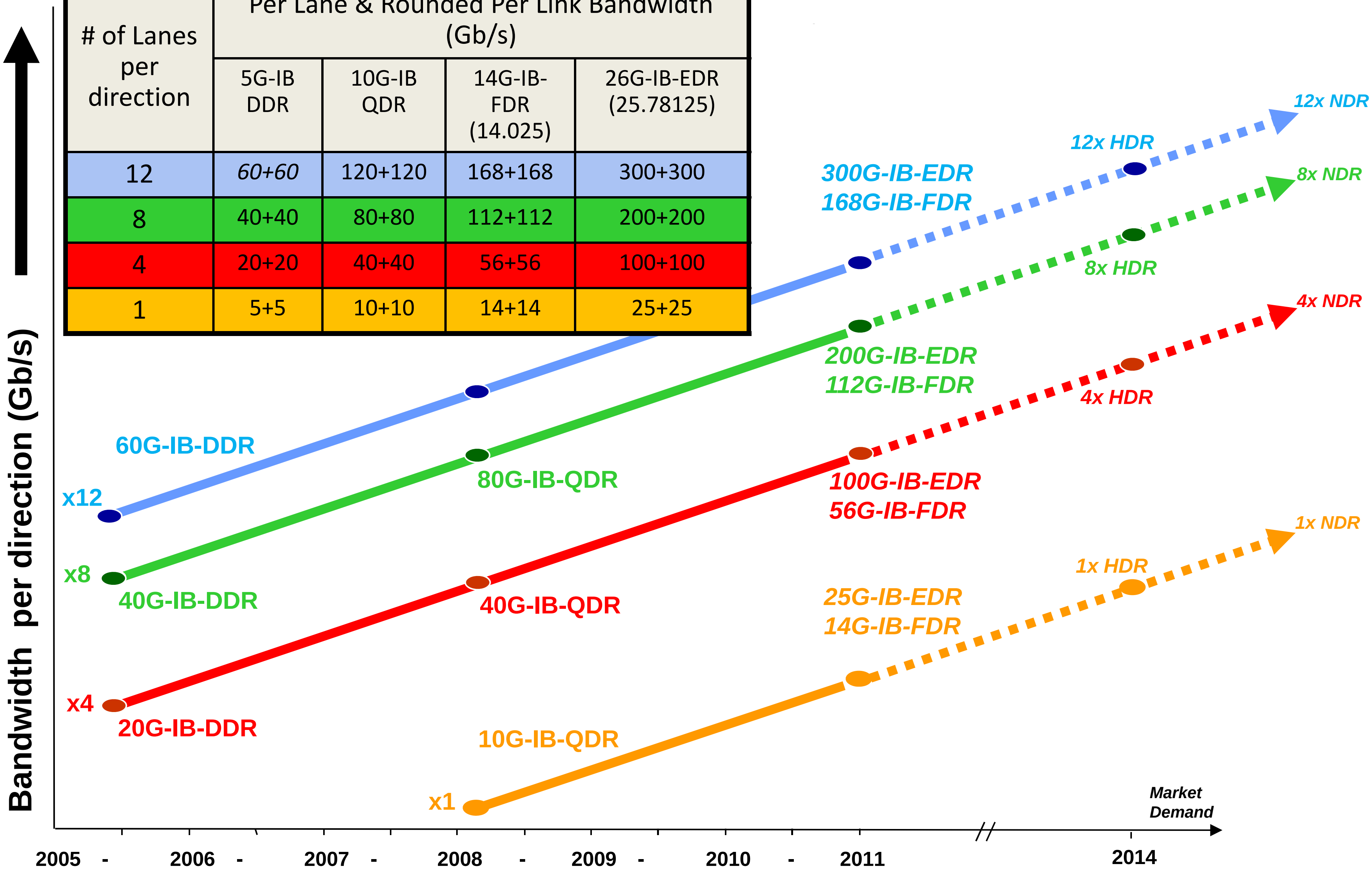
Thank You

Questions

Backup Charts

InfiniBand Link Speed Roadmap

# of Lanes per direction	Per Lane & Rounded Per Link Bandwidth (Gb/s)			
	5G-IB DDR	10G-IB QDR	14G-IB-FDR (14.025)	26G-IB-EDR (25.78125)
12	*60+60*	120+120	168+168	300+300
8	40+40	80+80	112+112	200+200
4	20+20	40+40	56+56	100+100
1	5+5	10+10	14+14	25+25



I/O Busses and Networks

- **PCI Express 2.0**
  - 1, 2, 4, 8, 12, 16, or 32 dual simplex 500 MB/s lanes (400 MB/s effective)
  - 8x = 4 GB/s (3.2 GB/s effective)
  - 16x = 8 GB/s (6.4 GB/s effective)
- **PCI Express 3.0**
  - Each lane is 1 GB/s (800 MB/s effective)
- **INFINIBAND**
  - 4x = 10 Gb/s DDR = 20 Gb/s QDR = 40 Gb/s FDR = 56 Gb/s
  - 8x = 20 Gb/s DDR = 40 Gb/s QDR = 80 Gb/s
  - 12x = 30 Gb/s DDR = 60 Gb/s QDR = 120 Gb/s
- **Fibre Channel**
  - FC4 = 400 MB/s
  - FC8 = 800 MB/s
  - FC16 = 1600 MB/s
- **Ethernet**
  - 1 Gb/s
  - 10 Gb/s
  - 40 Gb/s (4 x 10 GB/s per lane, QFSP)
  - 100 Gb/s (4 x 25 GB/s, QFSP)