

“The road to private cloud is paved  
with ethernet ports...”

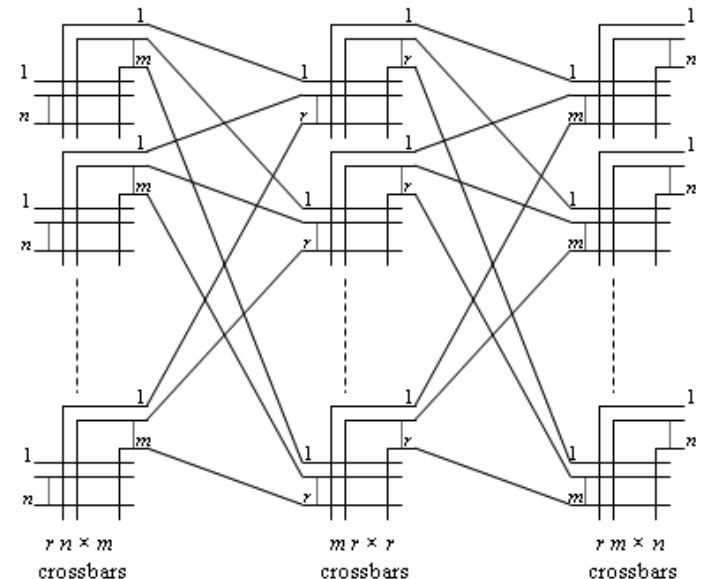
Luke Lonergan  
VP and CTO  
Data Computing Division

**UW MSR Summer Institute 2010**  
*Cloud Data Services: Challenges and Opportunities*

- Greenplum founded in 2003, now EMC
- Started deploying VMWare + Cisco + EMC one year ago
  - 1 Petabyte analytics system at T-Mobile
  - 10 Petabyte analytics system test for government
  - Key value proposition: elastic, agile, virtual
- The enabling technologies = GP MPP DB + virtualization + 10Gbit FCoE

# Bandwidth is cheap, latency is not

- A “Clos” network from Charles Clos in 1953 is a non-blocking switch built from tiers of crossbars
- Most modern switches are built internally as Clos or FAT Tree networks
- We build Clos networks by tiering switches from Cisco, Arista, Brocade, etc
- Bandwidth scales linearly to 1,200+ ports using two stages



- One of the principal benefits of virtualization is the ability to move services among physical servers
  - High Availability – mirror services
  - Load balancing on physical assets
  
- Relocation of state would require too much latency to be useful in these scenarios
  - The answer is SAN, but the old man SAN is too slow
  - Things are changing very quickly...

## What will we see next?



- Cisco has created an Ethernet standard for FCoE
  - They've been selling 10Gbit non-blocking networks for about 18 mos
  - Wide datacenter adoption
  - “Flattening out” of the datacenter network is part of it
- The impetus behind the datacenter re-work is cloud
  - Agility is key
  - Internal datacenters and CIOs are being compared with public cloud

## Why this matters to the Enterprise



- There are currently at least four networks managed by datacenters
  - Storage network on Fibre Channel
  - Application services on gigabit ethernet
  - System interconnects on Infiniband
  - Wide area networks on ATT, NTT, Verizon
- Which of these networks consumes the most admin time?
  - The FC SAN is a good candidate

## SAN – What is the current state-of-play?



- Locally attached disk drive (DAS)
  - 130 MB/s for a 15K RPM SAS disk
  - 100 MB/s for a 7.2K RPM SATA disk
  - Two host RAID controllers and 24 disks = 2,800MB/s
- Amazon EC2
  - 100 MB/s per (1Gbit) per EBS connection, in practice we see 80 MB/s
- FCoE attached SAN
  - 800 MB/s per FCoE link, in practice we see 700 MB/s
  - Two per host = 1,600 MB/s

- Current SSD technology landscape
  - SLC and MLC, one for enterprise, one for density
  - Biggest benefit is random access at  $\sim 10^3$  less latency than disk
  - Bandwidth in PCIe packaging is about 1.2 GB/s read, 1GB/s writes
  - Bandwidth in disk packaging is 280 MB/s read, 240 MB/s write
  - Limited by 3 Gbit SAS, will double next year
- Q1/2011, Intel will introduce all MLC line
  - With RAID and advanced wear leveling, MLC now durable enough for enterprise

- We're reaching the limit of SMP on the chip in 2-3 years
  - Somewhere in the neighborhood of 32 cores we should see the NUMA effects become important
- The answer for continued scaling is asymmetry
  - Two levels of programming, pipeline parallelism and SMP
- Memory->CPU bandwidth will continue to be a focus
  - Current practical limit is between 6-8 GB/s
  - Will need to grow like Moore's law to keep up

## Expected impact on programming methodologies



- Vectorized algorithms will be a big win
  - Most text analysis, scientific analysis and predictive models use matrix and vector computation
  - The database needs to adopt low level primitives that enable vectors (Martin's loving this right now)
- Optimizers need to get smart about asymmetry
  - And executors need to take advantage of the hierarchies in programming and memory
- Data and computation services must provide efficiency through smart relocation
  - Affinity, resource balancing, time sharing...

## White / black swans depending on your perspective



- Phase change memory
  - Recent events show that PCM may become a real player as soon as 2012
  - PCM is like memory in all but two areas: write *bandwidth* and durability
  - Price for capacity should be similar to SSD

**EMC<sup>2</sup>**<sup>®</sup>

**where information lives<sup>®</sup>**

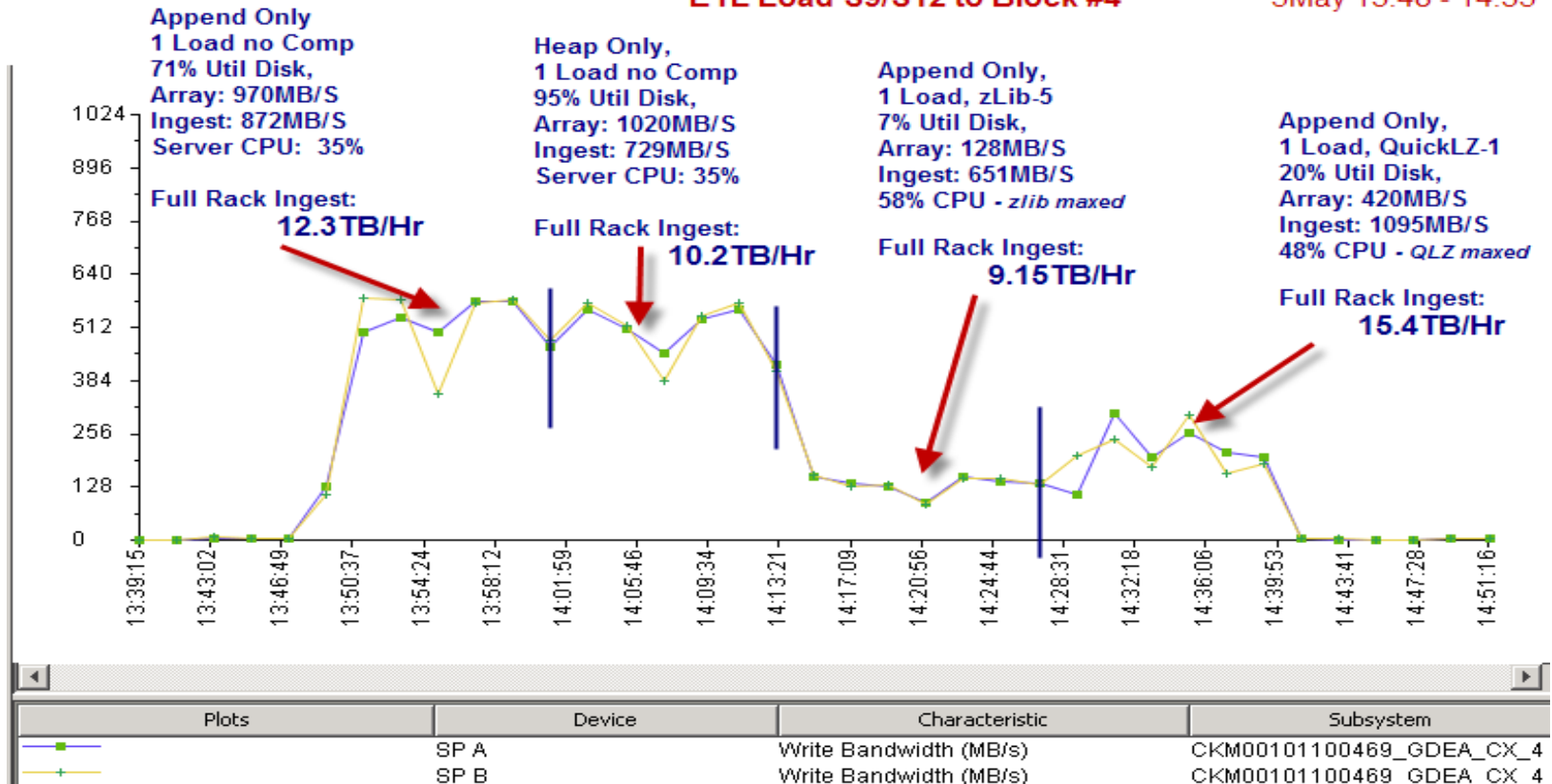
# Load/Ingest performance test results

## Row-level...



### ETL Load S9/S12 to Block #4

5May 13:48 - 14:35



- Fastest Load/Ingest with Greenplum's QuickLZ row-level compression
- 15.4TB/HR for full Rack solution - **3x times faster than Exadata**