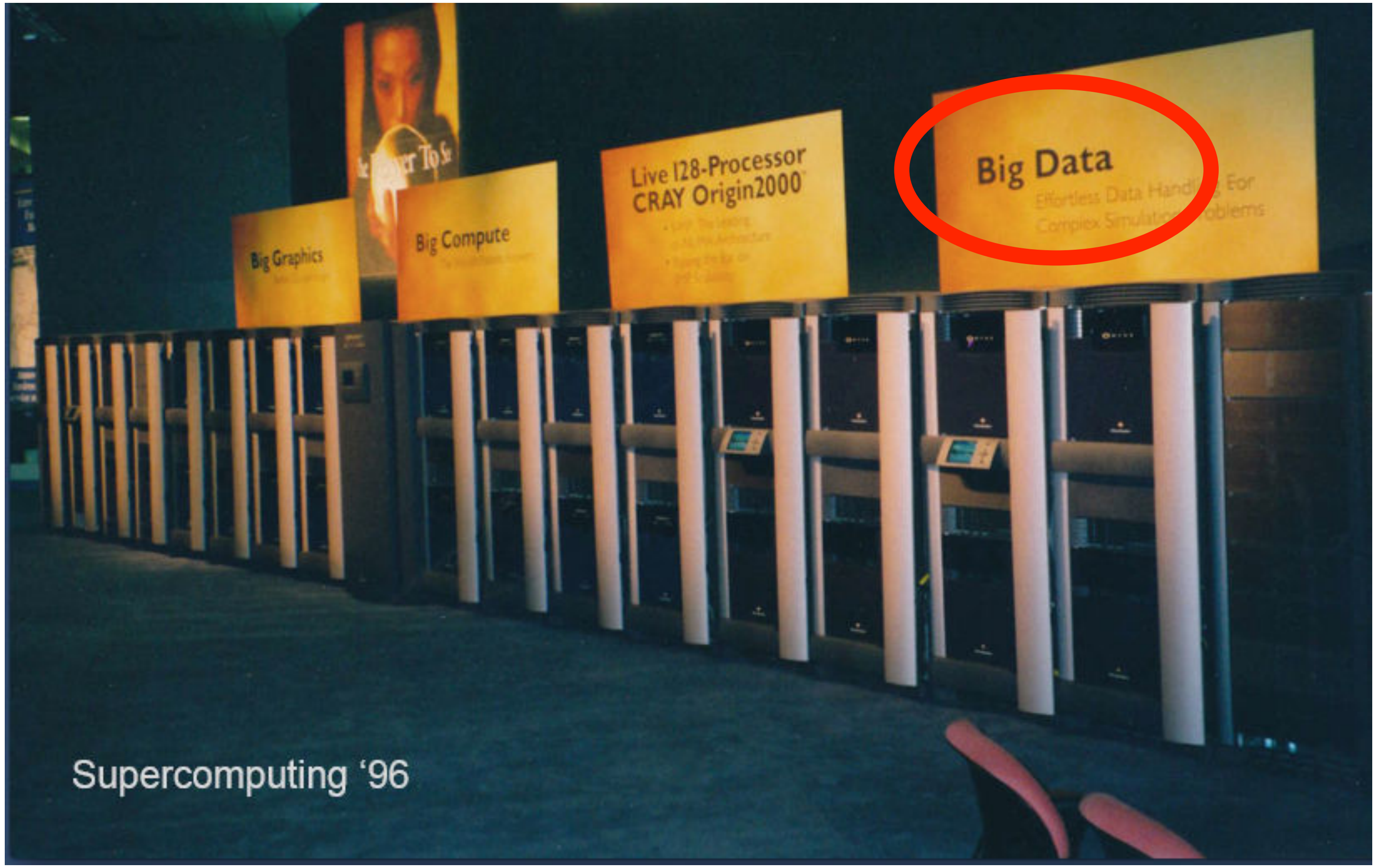
Data-Driven Science

SGI for Genomics Analysis

James Reaney, Ph.D., Senior Director, Research Markets

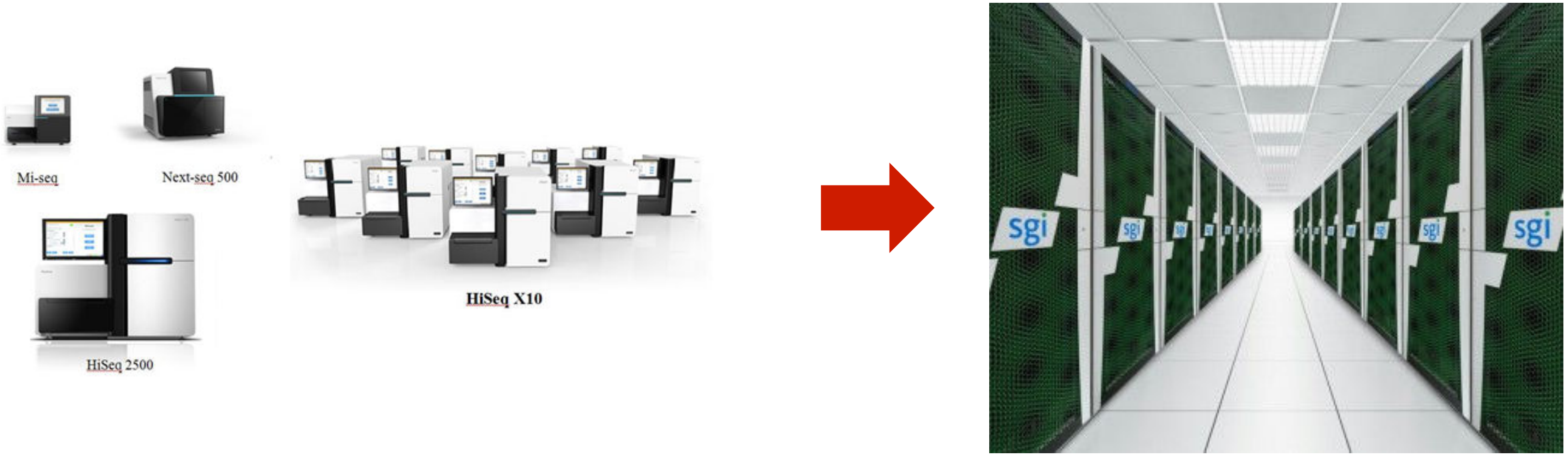
Did You Know?



Supercomputing ‘96

http://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/?_php=true&_type=blogs&_r=0

The New Normal



Actual quotes from Illumina execs

Mi-seq

Next-seq 500

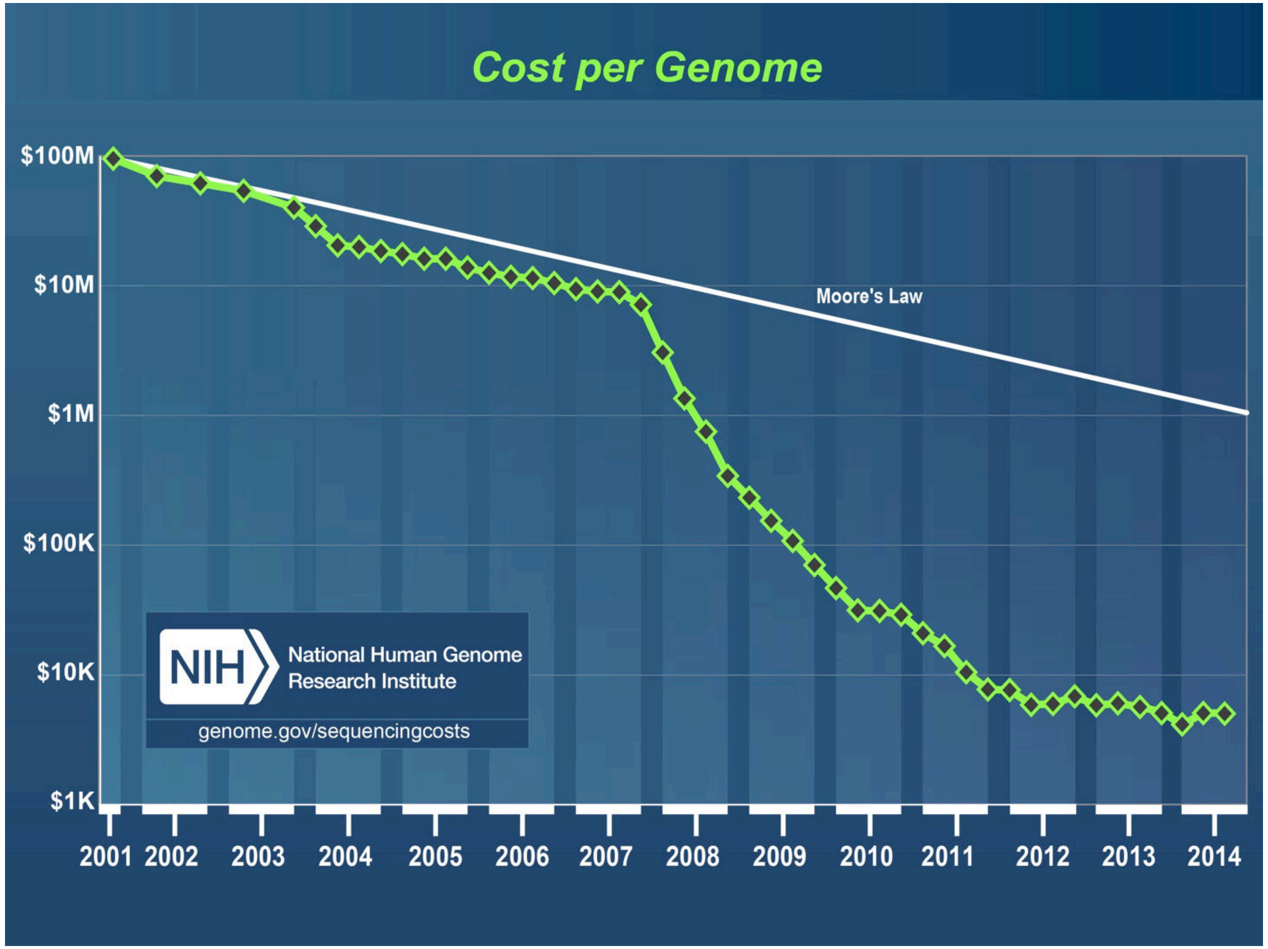
HiSeq 2500

HiSeq X10

- we are entering… “*the supersonic age of genomics*”
- the "demand for factory-scale sequencing of the human genome is *about to explode*“
  - Jay Flatley, CEO Illumina

- “Tens of thousands of samples are required…. You are trying to find needles in haystacks, and *you have to look at lots and lots of needles* to fundamentally understand the genetic basis of human disease“
- "Scientists are clearly finding clinical utility in the genome…. This creates a feedback loop where more discovery uncovers more clinical utility in the genome, which leads to an *increasing number of clinical researchers adopting these technologies*."
  - Christian Henry, SVP and CCO Illumina

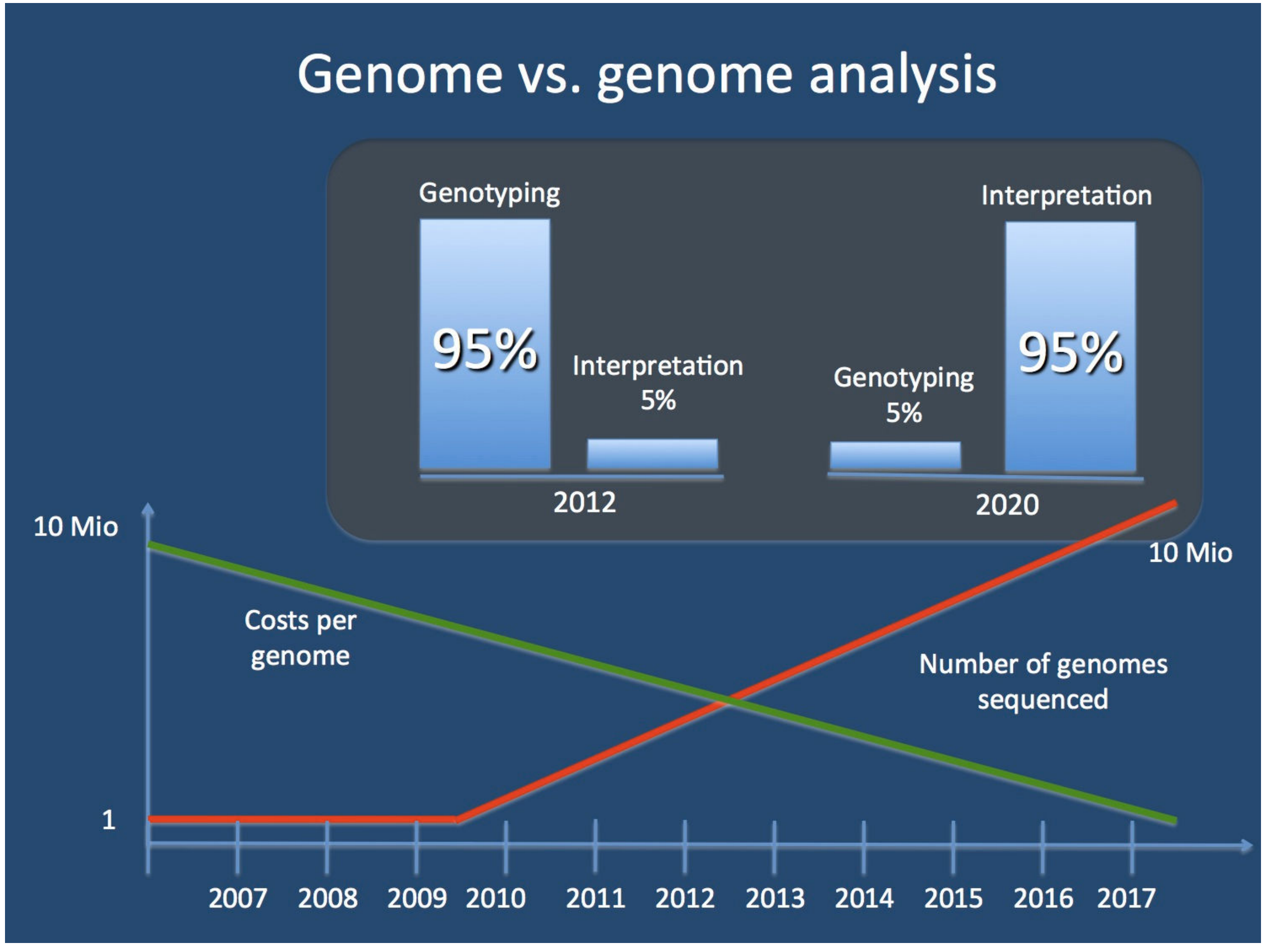
Democratization of DNA sequencing technology



- New sequencing tech enables far lower costs
- Lower costs drive availability of the research techniques to anyone, anywhere

- (*Psst*… it's really Kryder's Law we should be worried about anyway)

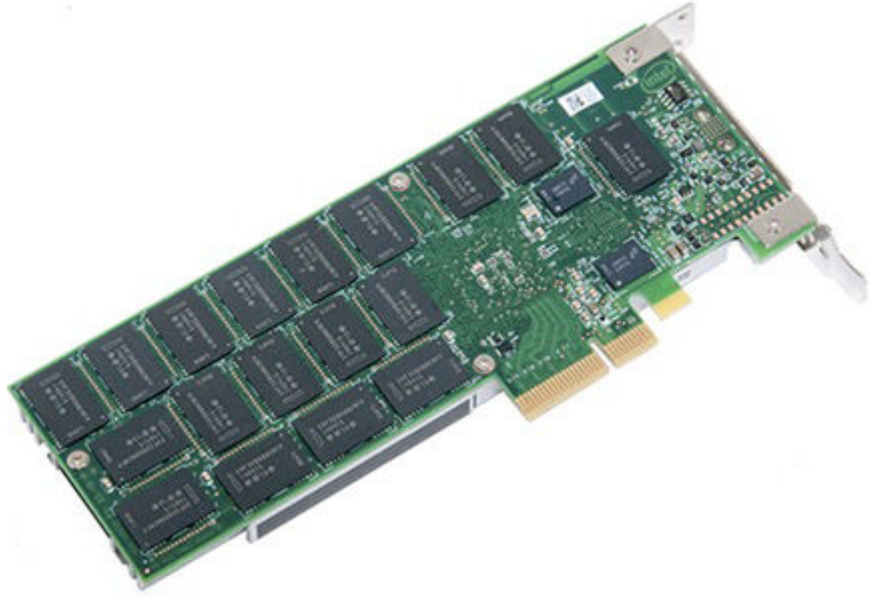
Now comes the hard part



- As sequencing costs fall, focus shifts to data analysis
- Want to guess what this means for storage architectures?
  - Parallel filesystems? No.
  - Hadoop? No.
- Need something better

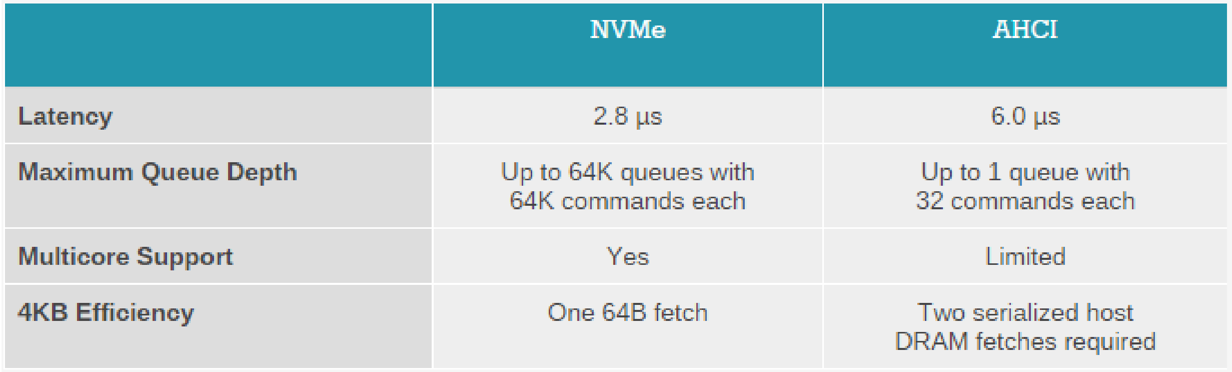
What is *NVMe* storage anyway?

- NVM = non-volatile memory, or “flash” memory, the same stuff in SSDs and your phone (only faster)
- The “e” stands for Express, because… well it just sounds ***FAST!***



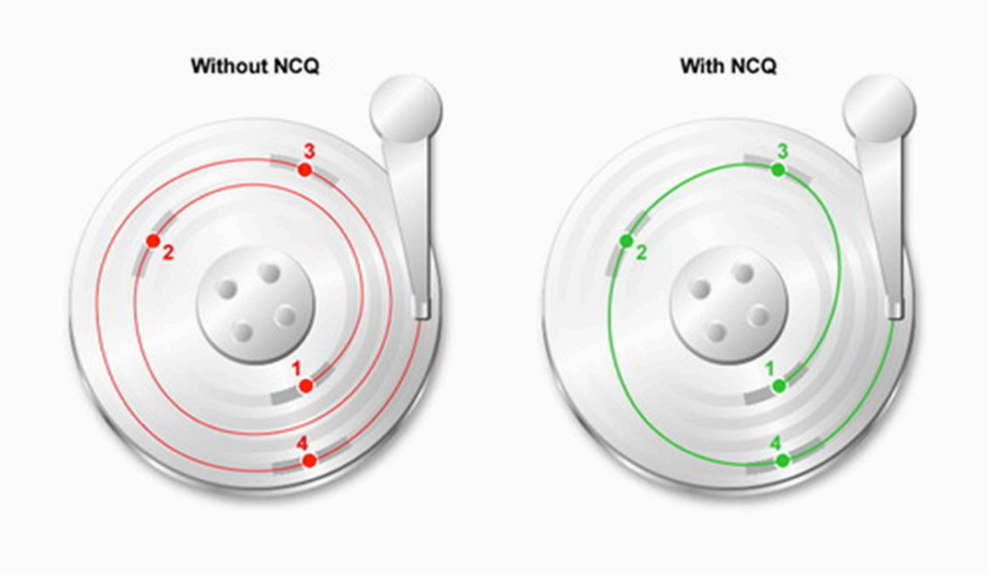
- Take SSD storage and stick on a PCIe card and this is what it looks like
- It’s an industry standards thing: Cisco, Dell, EMC, HGST, Intel, Micron, Microsoft, NetApp, Oracle, PMC, Samsung, SanDisk, and Seagate are all founding “promoter members” of NVMe

Why create *NVMe*?



	NVMe	AHCI
Latency	2.8 µs	6.0 µs
Maximum Queue Depth	Up to 64K queues with 64K commands each	Up to 1 queue with 32 commands each
Multicore Support	Yes	Limited
4KB Efficiency	One 64B fetch	Two serialized host DRAM fetches required

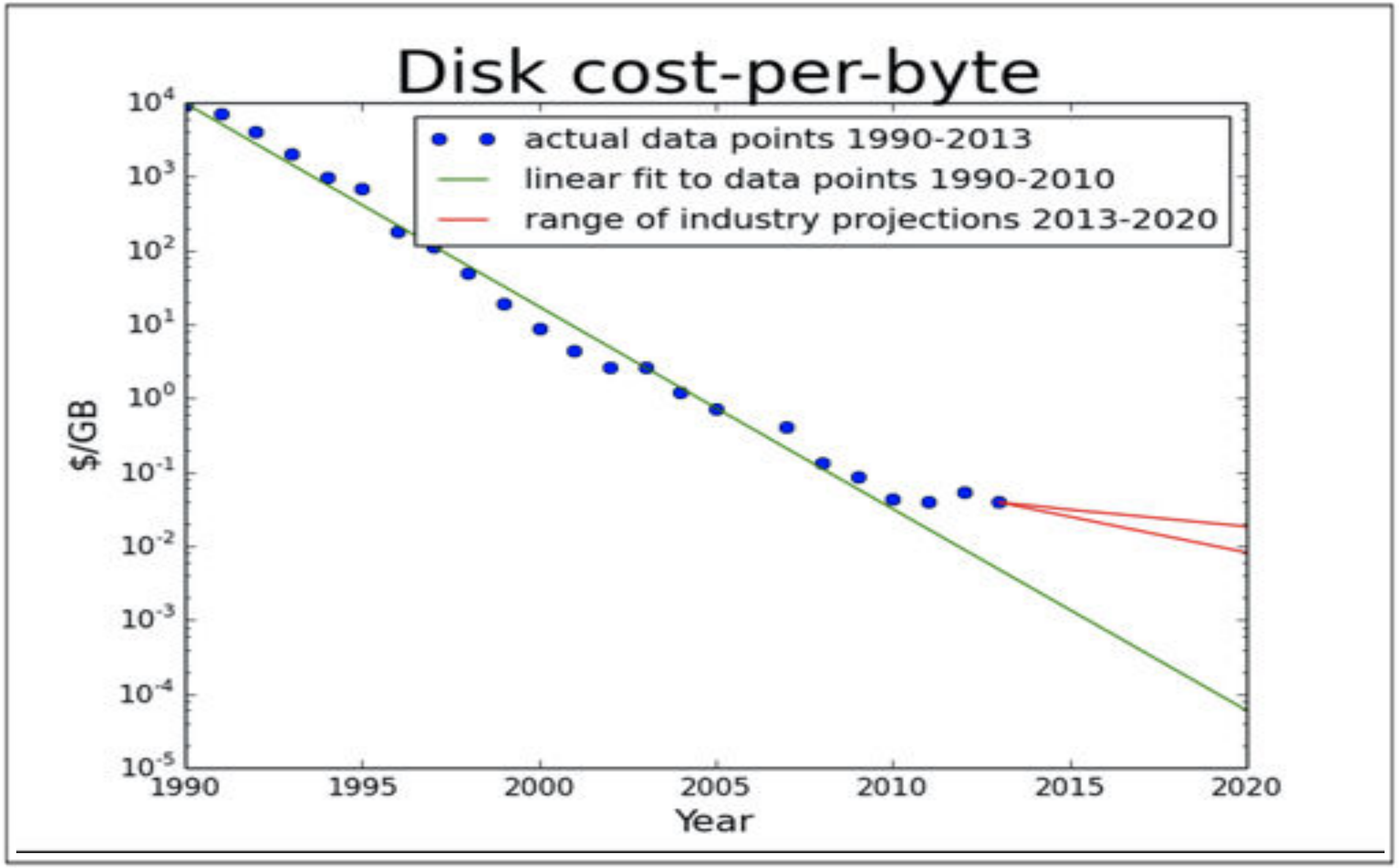
- Because the old storage standard, SATA (or more specifically AHCI) was not really created with fast storage in mind
- NVM behaves more like RAM than traditional spinning disk



- Too much latency: NVMe does away with kernel overhead, SATA/AHCI, SAS/SCSI, SAS expanders, FC or IB protocols, switch fabric, etc.

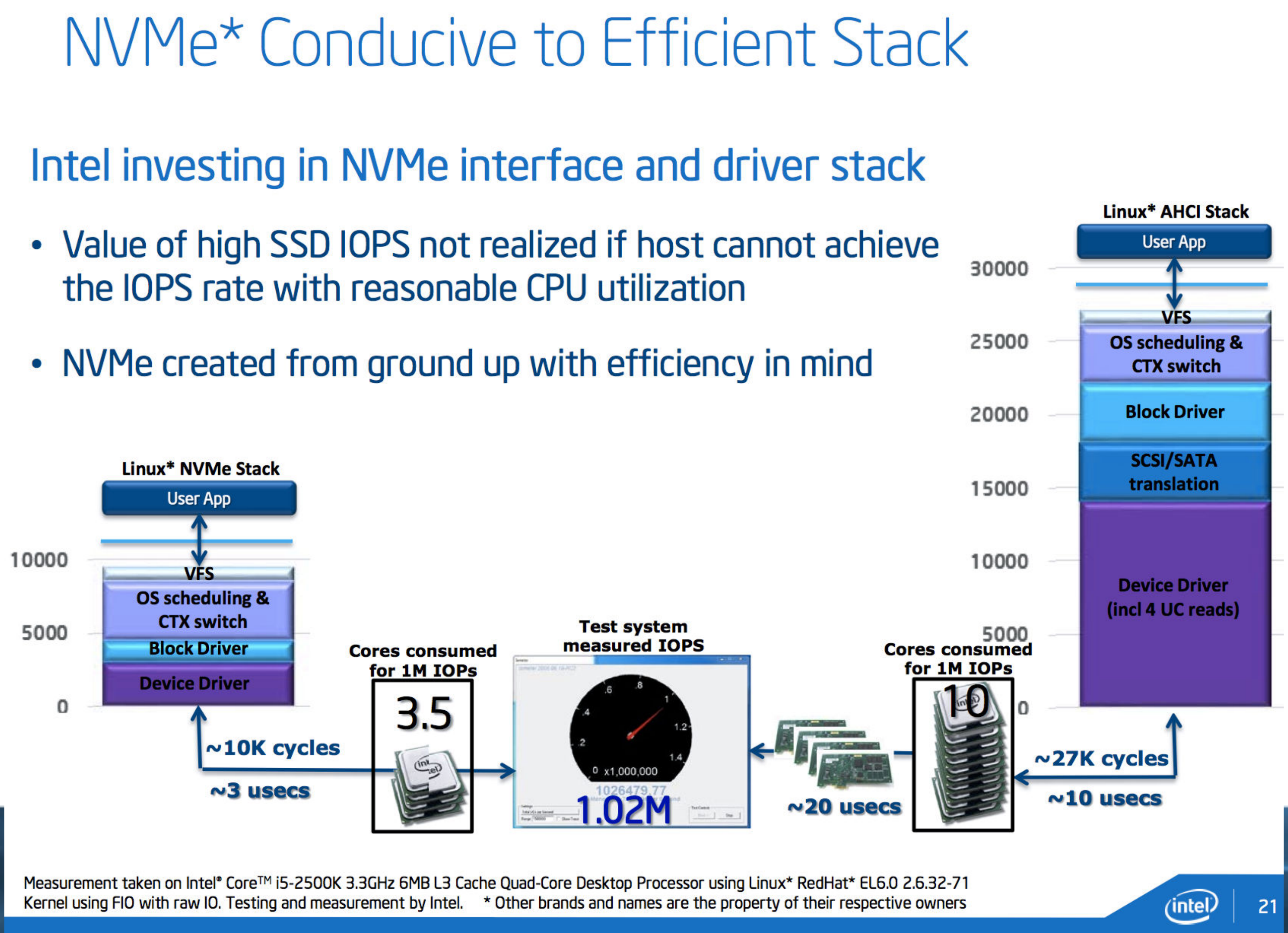
Moving the bottleneck

- What NVMe really does is get the fastest storage as close to the CPUs as possible, and that is what matters most
- Er… cheaper is always better too



Steven Foskett: http://blog.fosketts.net/2013/06/12/pure-ssd-fast/

(*I totally stole this slide from Intel®*)



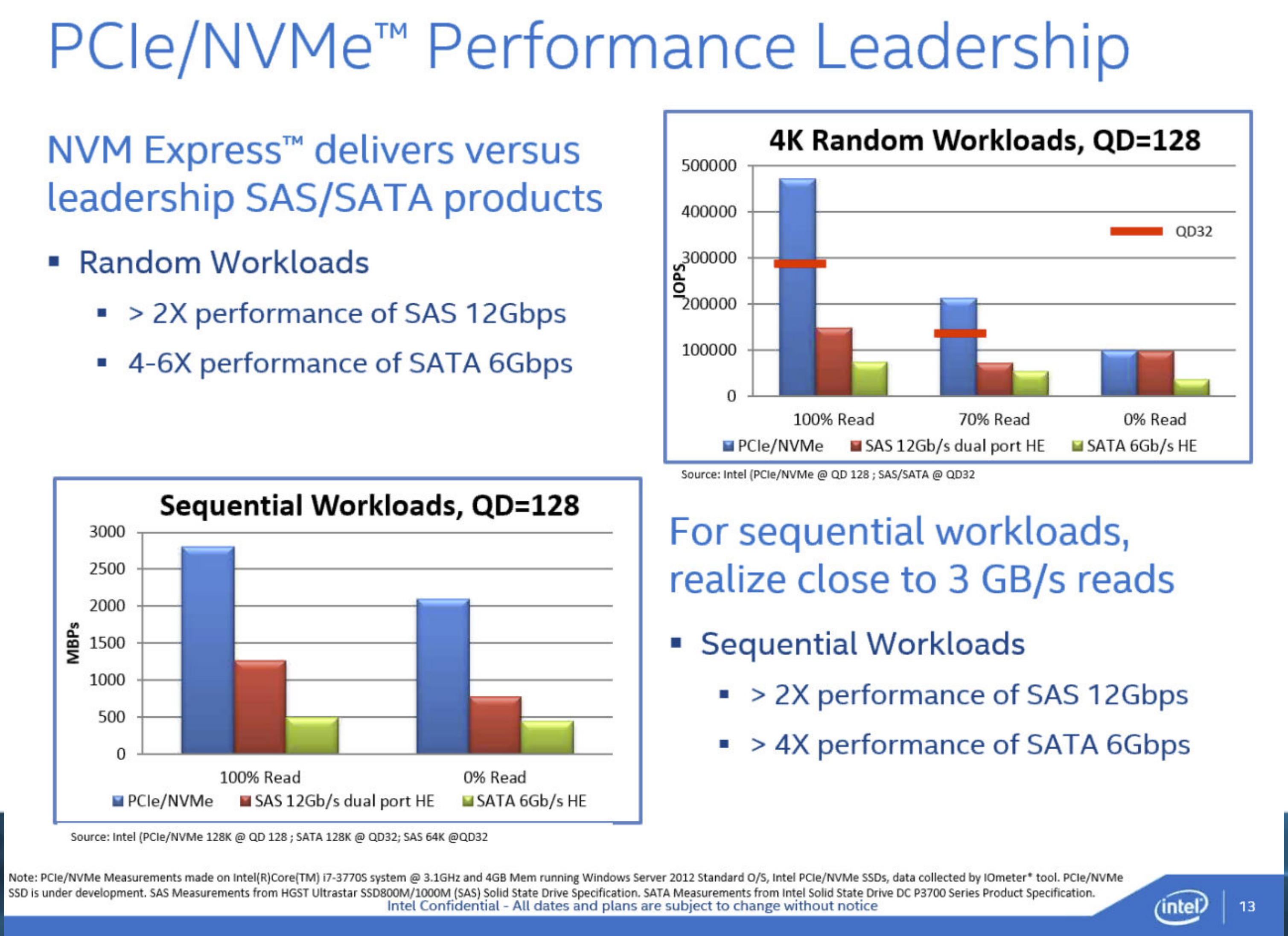
NVMe* Conducive to Efficient Stack

Intel investing in NVMe interface and driver stack

- Value of high SSD IOPS not realized if host cannot achieve the IOPS rate with reasonable CPU utilization
- NVMe created from ground up with efficiency in mind

Measurement taken on Intel® Core™ i5-2500K 3.3GHz 6MB L3 Cache Quad-Core Desktop Processor using Linux* RedHat* EL6.0 2.6.32-71 Kernel using FIO with raw IO. Testing and measurement by Intel. * Other brands and names are the property of their respective owners

(*I totally stole this slide from Intel® too*)



PCIe/NVMe™ Performance Leadership

NVM Express™ delivers versus leadership SAS/SATA products

- Random Workloads
  - > 2X performance of SAS 12Gbps
  - 4-6X performance of SATA 6Gbps

Source: Intel (PCIe/NVMe @ QD 128 ; SAS/SATA @ QD32

Source: Intel (PCIe/NVMe 128K @ QD 128 ; SATA 128K @ QD32; SAS 64K @QD32

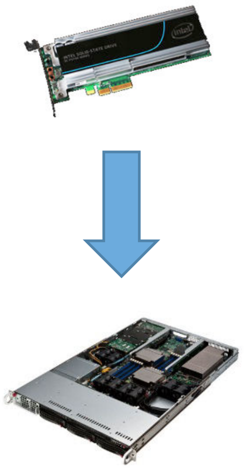
For sequential workloads, realize close to 3 GB/s reads

- Sequential Workloads
  - > 2X performance of SAS 12Gbps
  - > 4X performance of SATA 6Gbps

Note: PCIe/NVMe Measurements made on Intel(R)Core(TM) i7-3770S system @ 3.1GHz and 4GB Mem running Windows Server 2012 Standard O/S, Intel PCIe/NVMe SSDs, data collected by IOmeter* tool. PCIe/NVMe SSD is under development. SAS Measurements from HGST Ultrastar SSD800M/1000M (SAS) Solid State Drive Specification. SATA Measurements from Intel Solid State Drive DC P3700 Series Product Specification.

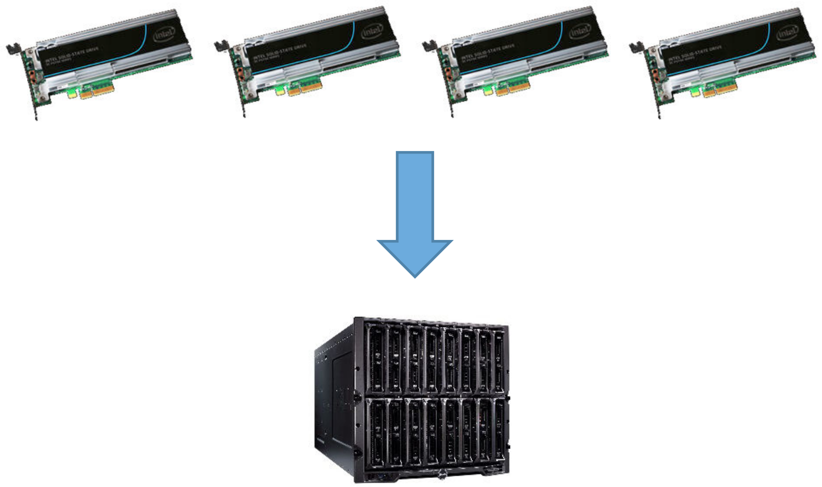
OK… faster is better, but how to best use NVMe?

- Remember, PCIe 3.0 spec is 8GT/s per lane (about 985MB/s), so a x4 PCIe bus has a theoretical maximum of a bit under 4GB/s



2TB raw capacity
Reads: 2.8GB/s Seq, 460k IOPS 4k Rdm
Writes: 2GB/s Seq, 175k IOPS 4k Rdm

How do I scale this?



32TB raw aggregate capacity ???
Reads: ???
Writes: ???

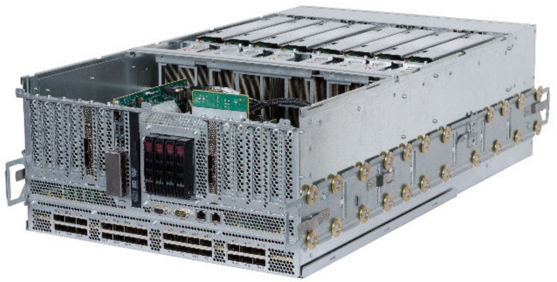
SGI UV 300 is how you do it

Overview

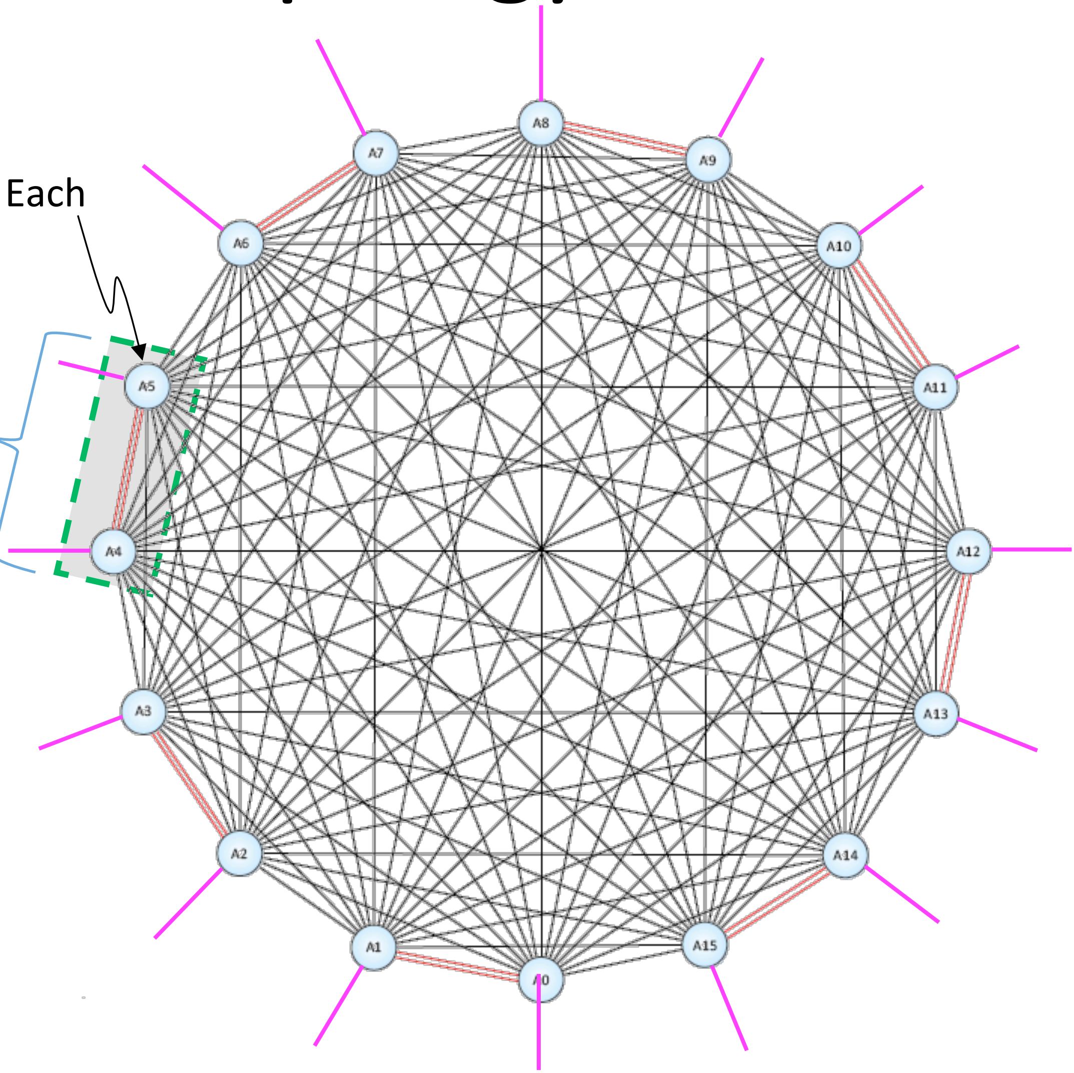
- UV 300 is the SGI's 7th generation, cache-coherent, shared-memory (NUMA) system based on Intel's Brickland (E7) processor family
- The scalable unit is a 5U chassis with 4 CPU sockets, memory, I/O and the SGI NUMALink™ interconnect (HARP2 ASIC)

HARP2 = 848GT/s (~105GB/s) @ 500ns
EDR IB (x4) = 100Gbs (~12.5GB/s) @ 500ns

- In one rack: 32-socket Single-System-Image
  24TB RAM using 32GB DIMMs
  96 PCIe 3.0 slots (32 x16 and 64 x8)
- Max scale: 2048 cores/threads
  48TB RAM using 32GB DIMMs
  192 PCIe 3.0 slots



SGI UV 300 Topology



2 Sockets Each

UV300 Chassis
4 CPU Sockets
96 DIMMs
12 PCIe Slots
(typical)

Up to:

32 sockets, All-to-All topology
24TB Memory (32GB DIMMs)
96 PCIe Gen3 slots total
* 32 PCIe x16 Gen3 slots
* 64 PCIe x8 Gen3 slots
Fits in one 42U Rack

Intel® QuickPath ═

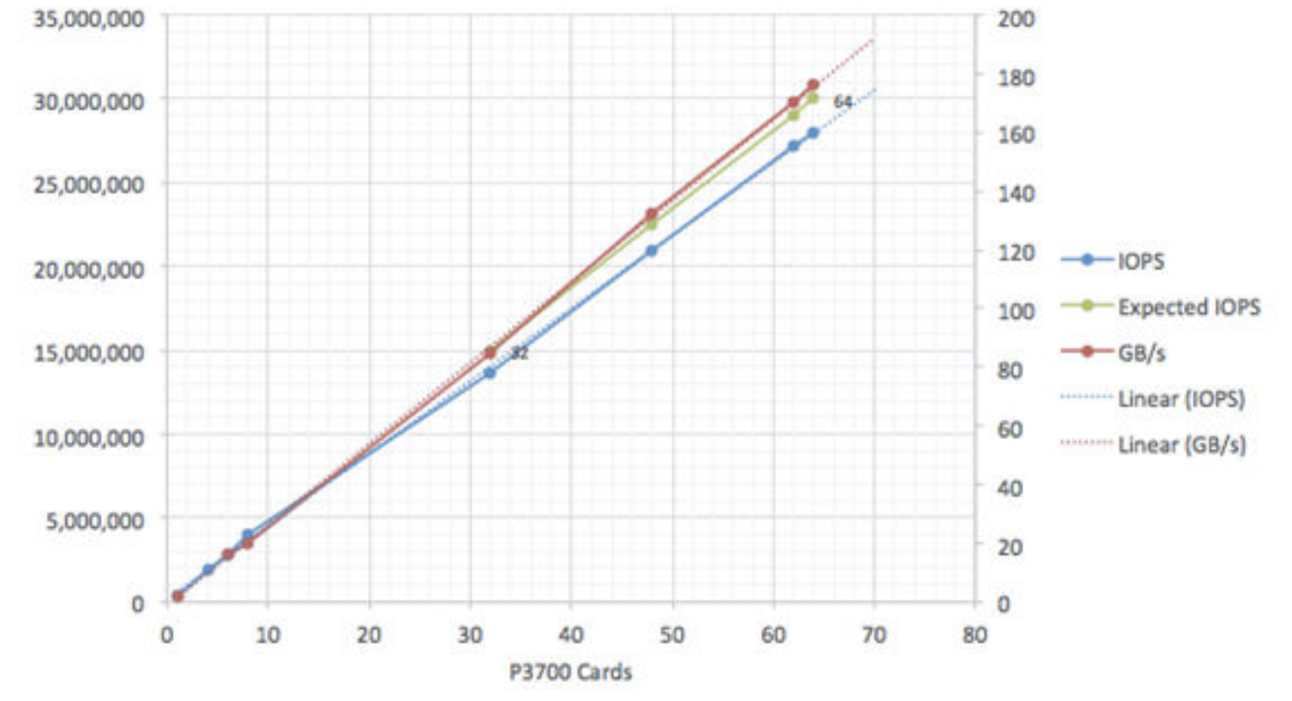
SGI® NUMAlink —

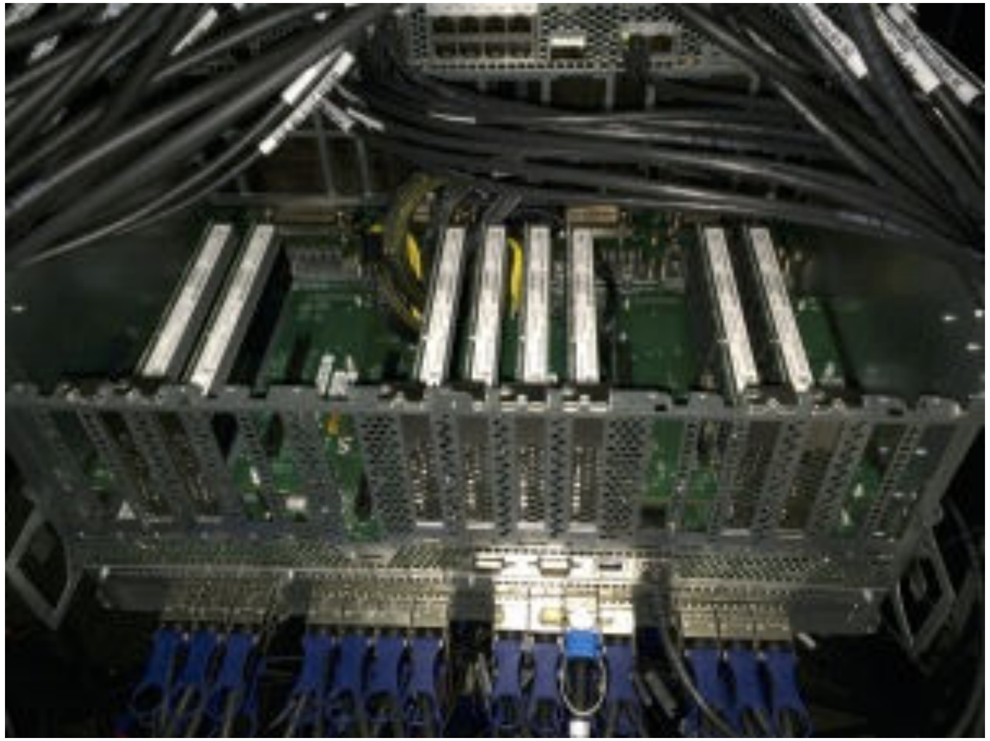
2 CPU Sockets ○

Each line = —
6x PCIe slots
2x x16
4x x8

“SGI racks UV brains, reaches *30 MEEELLION IOPS”*

- Eight UV300 chassis in one rack, 32 socket system, 24TB RAM, 64 NVMe cards (8 per chassis)





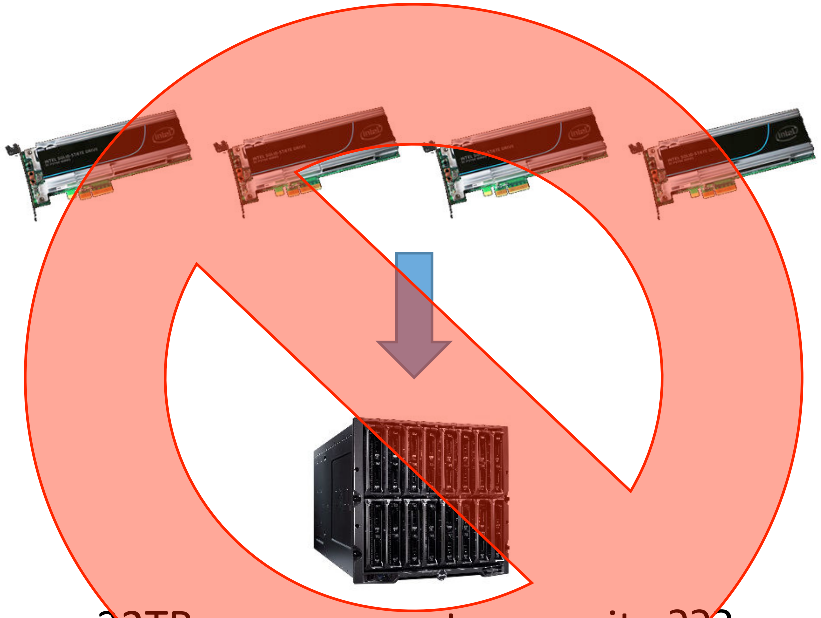
Chris Mellor: http://www.theregister.co.uk/2014/11/17/sgi_uv_reaching_30_million_iops_with_nvme_flashers/

SGI UV: the capability platform with *ridiculous* I/O

- Future-proofing = not worrying about hitting bottlenecks

192TB raw capacity
240GB/s
44 million IOPS

Scale like this!



32TB raw aggregate capacity ???
Reads: ???
Writes: ???

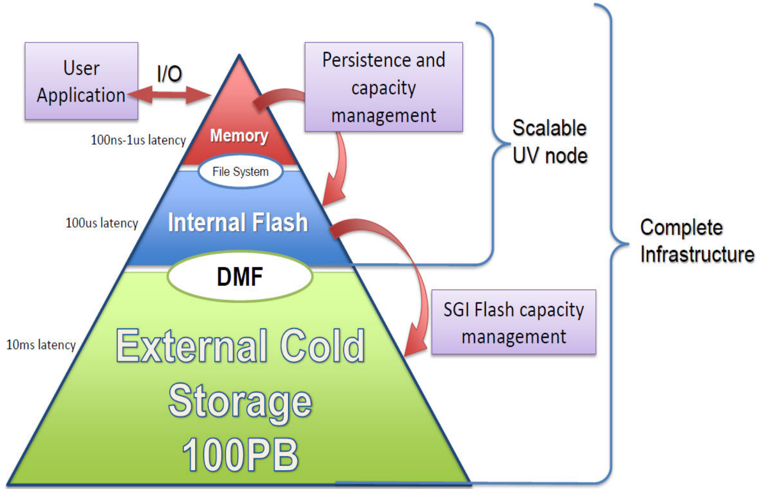
SGI UV: but wait, there’s (a lot) more

- DMF = Data Migration Facility

192TB raw capacity
240GB/s
44 million IOPS

+

=



Applicability to genomics workflows

- SDSC, using Gordon to investigate whole-genome sequencing of 438 patients for rheumatoid arthritis
  - 50TB aggregate DNA sequences, 350TB peak project storage, only ~5k cores
  - 14-stage genomics pipeline, I/O-bound vs. compute-bound
  - 6 weeks, 300k hours on Gordon instead of 4 years

"…big data challenges such as human genomics would dictate new supercomputer architectures where memory and IOPS (I/O operations per second) would be more important than raw computing power…."

-- Michael Norman, SDSC Director

http://ucsdnews.ucsd.edu/pressrelease/sdsc_assists_in_whole_genome_sequencing_analysis_under_collaboration_with_j

Applicability to genomics workflows

“Computers are the new microscopes, and data is the new blood draw”

-- Rajesh Gupta, department chair, CS&E UCSD

- Franz Och (Google Translate) joins Human Longevity (J Craig Venter)

- Jill Mesirov (Broad Institute) joins UCSD School of Medicine
- Rob Knight (University of Colorado) joins UCSD School of Medicine

“I want us to lead the field in precision medicine, and computational biology is part of that… we need to be able to handle large data sets.”

-- David Brenner, dean, UCSD School of Medicine

http://m.utsandiego.com/news/2015/apr/15/bigdata-stars-ucsd/

We don’t know what we don’t know

- SARS, sequenced in 31 days (way back in 2003)
- Ebola, genomic “surveillance” used to characterize viral transmission patterns
- MRSA, NGS used to identify and track spread of disease

- When the next black swan event hits, we will have the means to extract relevant scientific knowledge, quickly

SGI: Solve the Big Problems