

# Trends in Scientific Discovery Engines

Mark Stalzer

Center for Advanced Computing Research  
California Institute of Technology

[stalzer@caltech.edu](mailto:stalzer@caltech.edu)  
[www.cacr.caltech.edu](http://www.cacr.caltech.edu)

January 31, 2013







*This is an informal talk about trends in the engines that make scientific discovery happen faster...*

*and things are happening much faster than expected and much slower than we think...*

*follow the **POWER...***

*(The working group will cover broader cyber-infrastructure issues)*



# Drivers

- Simulations
  - ▶ Benchmark: LINPACK Benchmark (TOP500)
- Big Data
  - ▶ Benchmark: ? (*This is a problem*)
- Power
  - ▶ Tflops/KW (for Rmax): 2.14 #1, 2.07 Sequoia #2, 2.46 best  
Programability: Sequoia, Beacon, #1





*Trends in Simulation Engines:*


*Exascale or Bust*

*(or computing without data)*



# Top 10 Supercomputers

## June 2012

Rank	Site	Computer/Year Vendor	Cores	$P_{peak}$	$P_{sust}$	Power
1	DOE/NSA/LLNL United States	Sequoia - BlueGene/Q, Power 80C 16C 1.60 GHz, Custom / 2011 IBM	1572864	16324.75	20132.6	7890.0
						
2	Germany	IBM	131072	1360.09	1677.04	680.0
3	CEA/TOCC-GENCI France	Curie thin nodes - Bullx B010, Xeon ES- 2880 8C 2.700GHz, Infiniband QDR / 2012 Bull	77184	1359.00	1667.17	2251.0
10	National Supercomputing Centre in Shenzhen (NSCC) China	Nebulae - Dawning TC3600 Blade System, Xeon X3650 8C 2.80GHz, Infiniband QDR, NVIDIA 2050 / 2010 Dawning	120640	1271.00	2984.30	2580.0



# Socket Parallelism: TOP500 2002 vs. 2012



Optimized for dot products:

```
complex s = 0;
```

```
for (complex* lp = xp + n; s += *xp++ * *yp++; xp < lp) ;
```

	ASCI White #2	Sequoia #1	GAIN
Rmax (Tflops)	7.226	16,325	2,260x
Processor	IBM Power 3	IBM BQC 16C	
Clock (Ghz)	0.375	1.6	4.27x
#Sockets	8,192	98,304	12x
Socket Parallelism	1	64	<b>64x</b> <b>(3,280)</b>

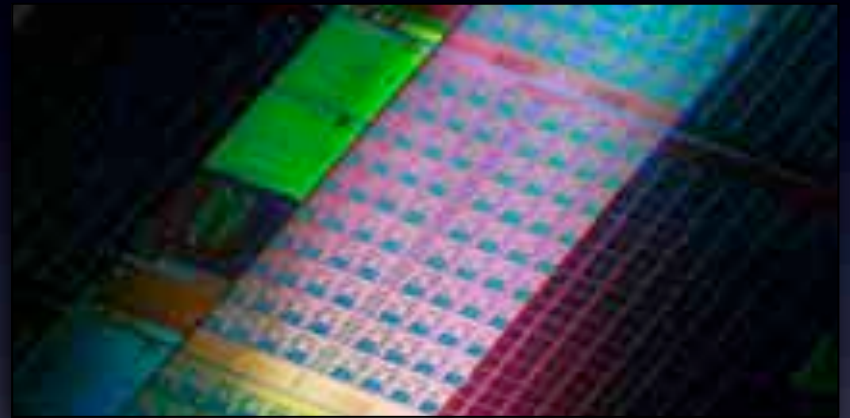
*Most of parallelism increase is on-socket, and getting “worse”...*



# Intel Xeon Phi

November 2012

- 50+ IA64 cores (Pentium-like)
  - ▶ 8 (double) SIMD
  - ▶ 200+ hardware threads
  - ▶ Tflops per socket
- Will scale to  $O(1,000)$  threads ~3 yrs
  - ▶ Cache coherent in socket
  - ▶ ASCI White performance!
- Qualitatively different programming model
  - ▶ MPI between sockets
  - ▶ Massive **general** threading in a socket: must rewrite codes
- *Example*: Image processing parallelism



“Socket Archipelago”

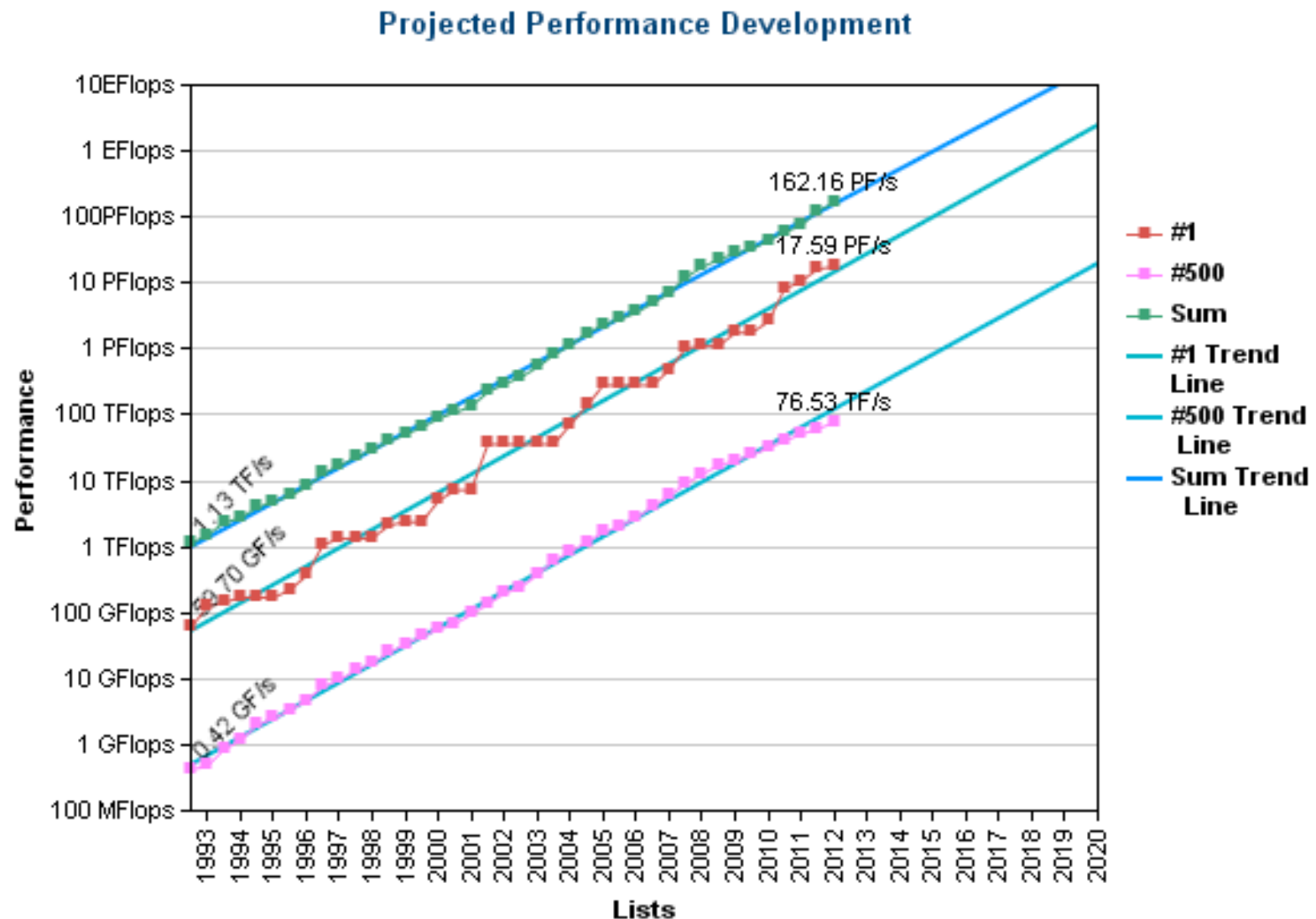


# Beacon: Most Energy Efficient

Beacon - Appro GreenBlade GB824M, Xeon E5-2670 BC 2.600GHz, Infiniband FDR, Intel Xeon Phi 5110P							
Site:	National Institute for Computational Sciences/University of Tennessee						
Manufacturer:	Appro International						
Cores:	9216						
Linpack Performance (Rmax)	110.5 TFlop/s						
Theoretical Peak (Rpeak)	157.5 TFlop/s						
Power:	45.11 kW						
Memory:	9216 GB						
Interconnect:	Infiniband FDR						
Operating System:	Linux						
Compiler:	Intel Compiler						
Math Library:							
MPI:	Intel MPI						
Ranking							
List	Rank	System	Vendor	Total Cores	Rmax (TFlops)	Rpeak (TFlops)	Power (kW)
11/2012	1	Appro GreenBlade GB824M, Xeon E5-2670 BC 2.600GHz, Infiniband FDR, Intel Xeon Phi 5110P	Appro International	9216	110.5	157.5	45.11



# Top 500 Trends





# Current Exascale Construction: Intel Fab 42 (14nm and beyond)





2018

# UQ +Solvers and Extreme Parallelism

25MW

POWER

UQ INTRINSIC PARALLELISM

UQ Optimizer

Convergence Studies

many iterations

"few" runs

10 per iteration

Solvers

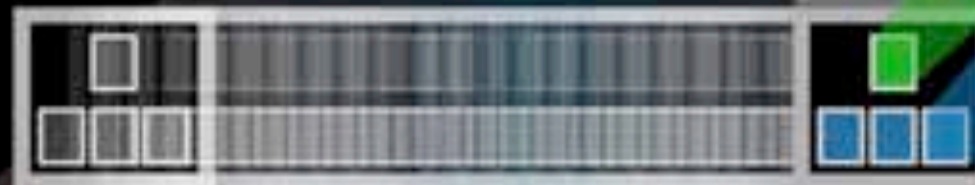
PROCESSOR + SHARDING

10,000 way job

NODES + MPI



SOCKET<sup>®</sup> +  
THREADING



(Socket Processor)  
Deep Pipelines

250<sup>+</sup> cores  
1,000<sup>+</sup> threads  
Shallow Pipelines

std

Single Instruction  
Multiple Data  
1024 bits

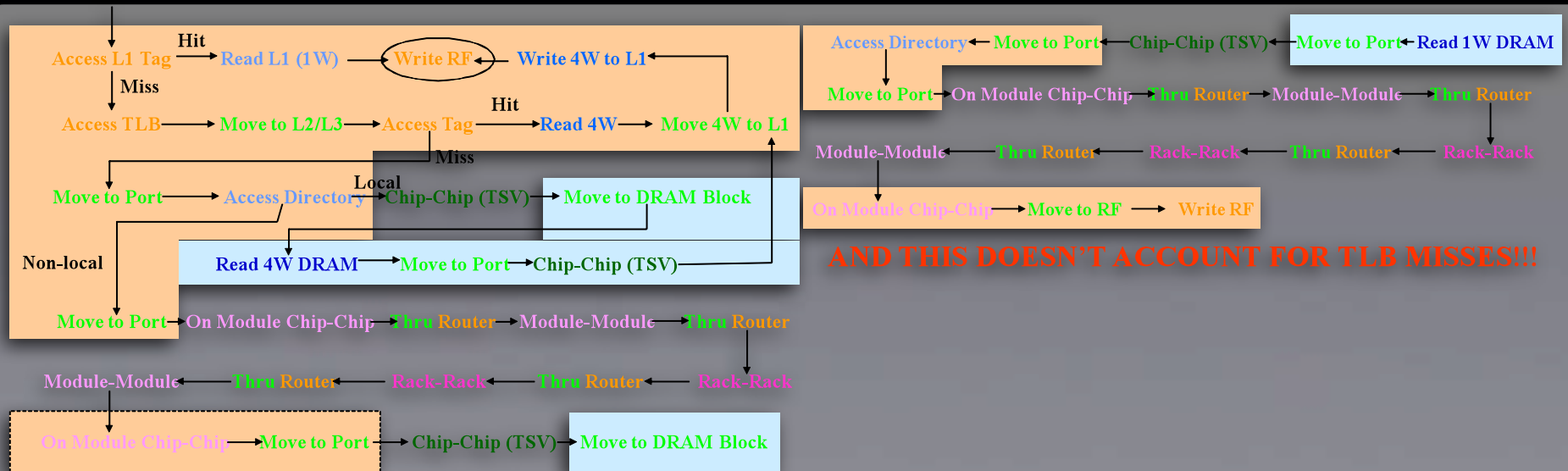
$10^{18}$  flops

\* 100,000 Socket Constraint





# Follow the Power



<u>Operation</u>	<u>Energy (pJ/bit)</u>
Register File Access	0.16
SRAM Access	0.23
DRAM Access	1
On-chip movement	0.0187
Thru Silicon Vias (TSV)	0.011
Chip-to-Board	2
Chip-to-optical	10
Router on-chip	2

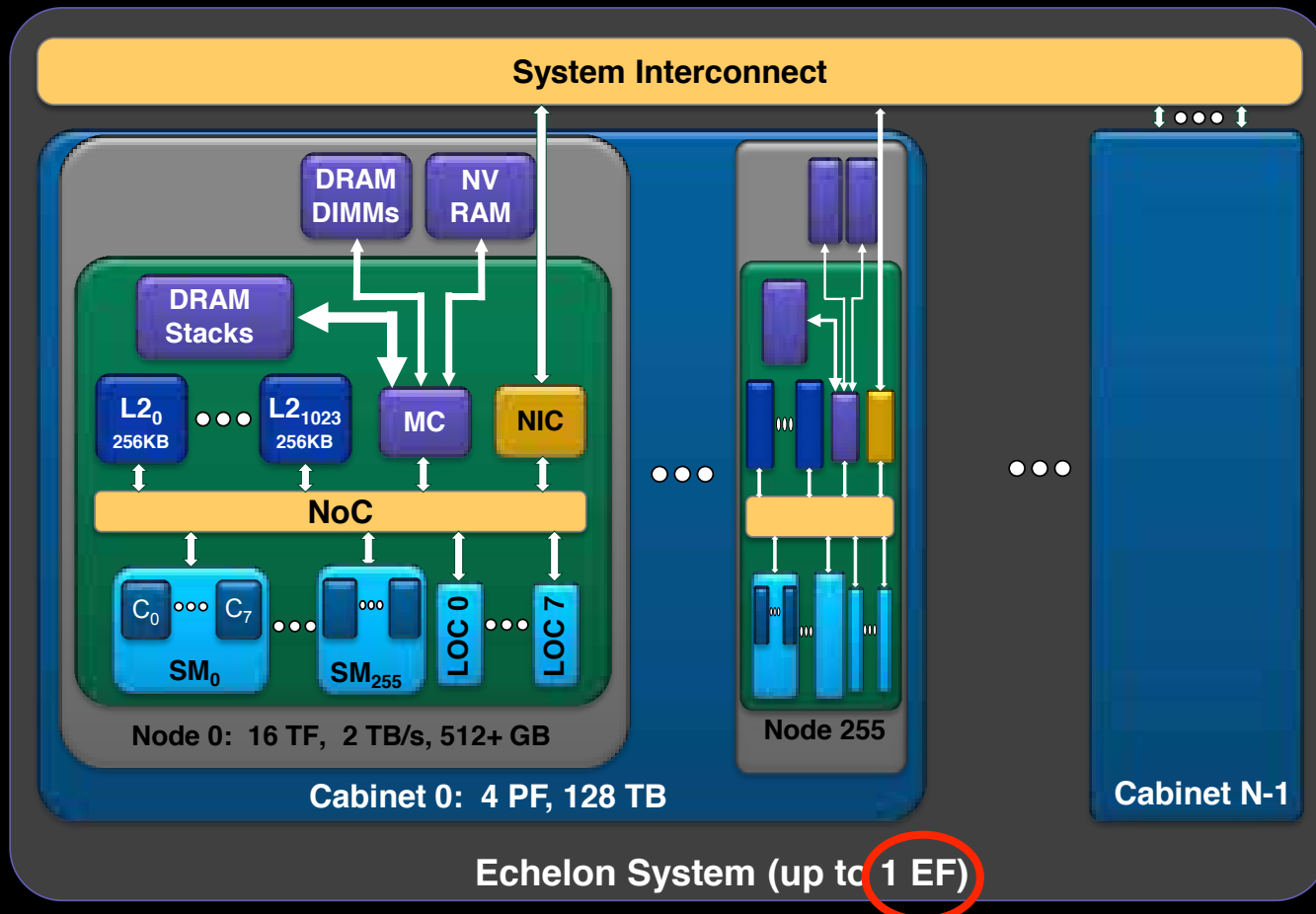
Step	Target	pJ	#Occurrences	Total pJ	% of Total
Read Alphas	Remote	13,819	4	55,276	16.5%
Read pivot row	Remote	13,819	4	55,276	16.5%
Read 1st Y[i]	Local	1,380	88	121,400	36.3%
Read Other Y[i]s	L1	39	264	10,425	3.1%
Write Y's	L1	39	352	13,900	4.2%
Flush Y's	Local	891	88	78,380	23.4%
Total				334,656	
Ave per Flop				475	

**In 2015, a flop will be ~10 pJ: It takes ~50x energy just to move the bits!**



# NVIDIA Echelon

## 2018 Echelon Compute Node & System



Key architectural features:

- Malleable memory hierarchy
- Hierarchical register files
- Hierarchical thread scheduling
- Place coherency/consistency
- Temporal SIMT & scalarization
- PGAS memory
- HW accelerated queues
- Active messages
- AMOs everywhere
- Collective engines
- Streamlined LOC/TOC interaction

Copyright 2012, NVIDIA

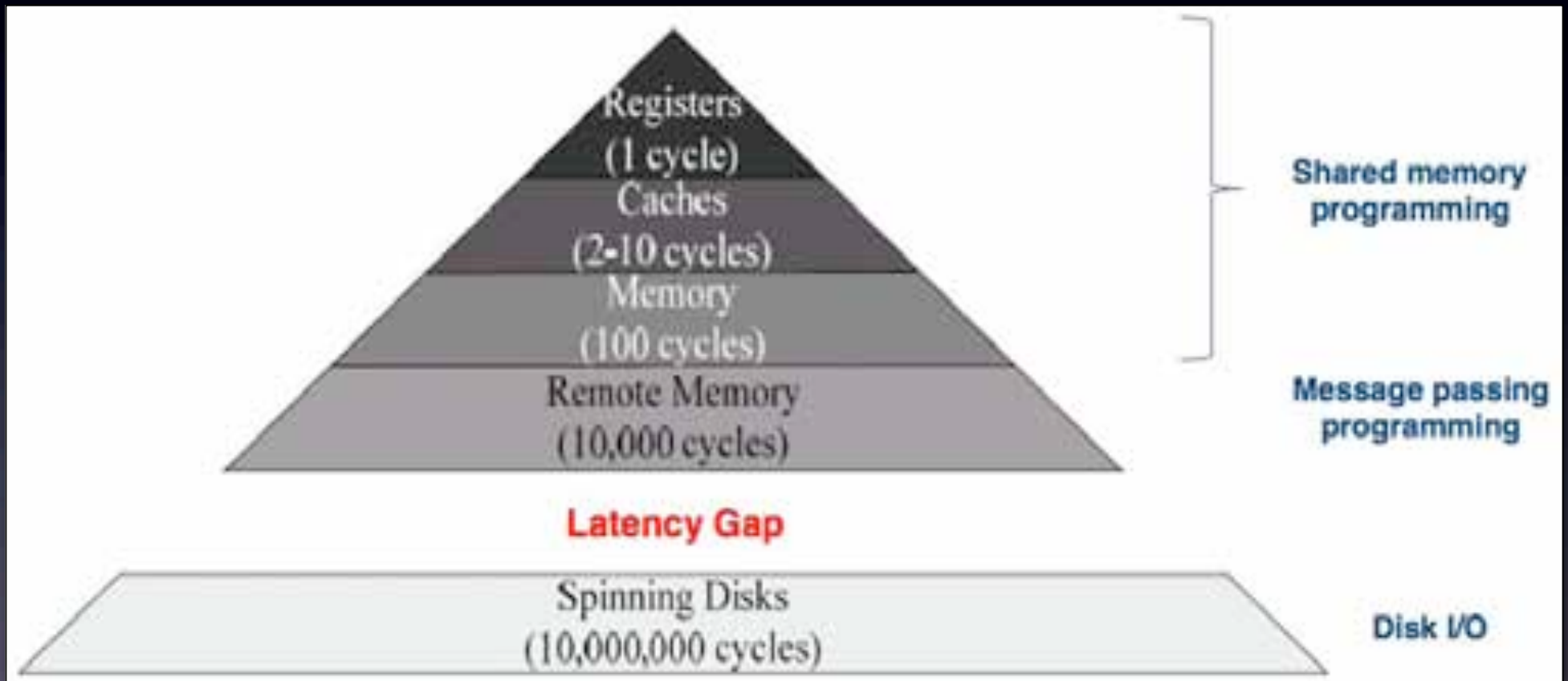


## *Trends in Data to Discovery Engines*

*(Computing with data, because the answer is not always 42)*



# Latency Gap



*Several efforts at closing the latency gap using flash memories...*



# Amdahl-Balanced Blades



- Gene Amdahl's Laws for I/O & memory (1965, 2007):
  - ▶ A bit of seq. I/O per sec. per instruction per sec. (*Amdahl #*)
  - ▶ Mbytes / MIPS  $\sim 1$  (*Memory ratio*)
  - ▶ One I/O operations per 50,000 instructions (*IOPS ratio*)
- Simulation codes may have an Amdahl # of  $10^{-5}$ ; data intensive apps may need  $\sim 1$
- Szalay, Bell, Huang, Terzis, White (Hotpower-09):

Table 2: Performance, power, and cost characteristics of various data-intensive architectures.

	CPU [GHz]	Mem [GB]	SeqIO [GB/s]	RandIO [kIOPS]	Disk [TB]	Power [W]	Cost [\$]	Relative Power	Amdahl numbers		
									Seq	Mem	Rand
GrayWulf	21.3	24	1.500	6.0	22.5	1,150	19,253	1.000	0.56	1.13	0.014
ASUS	1.6	2	0.124	4.6	0.25	19	820	0.017	0.62	1.25	0.144
Intel	3.2	2	0.500	10.4	0.50	28	1,177	0.024	1.25	0.63	0.156
Zotac	3.2	4	0.500	10.4	0.50	30	1,189	0.026	1.25	1.25	0.163
AxiomTek	1.6	2	0.120	4.0	0.25	15	995	0.013	0.60	1.25	0.125
Alix 3C2	0.5	0.5	0.025	N/A	0.008	4	225	0.003	0.40	1.00	



# Cyberbricks

- 36-node Amdahl cluster at 1,200 W total!
  - ▶ N330 dual core Atom, 16 GPU cores, 4 GB
- Aggregate disk space of ~43 TB
  - ▶ About 1 SSD 120 GB per core (~8 TB)
  - ▶ 35 TB of spinning disk
- Blazing I/O performance: 18 GB/s
- Amdahl # = 1 for under \$30 K
- Using the GPUs for data mining:
  - ▶ 6.4 B multidimensional regressions in 5 min over 1.2 TB
  - ▶ Ported RF module from R to C#/CUDA





# Gordon Supernode Architecture

## 32 Appro Extreme-X compute nodes

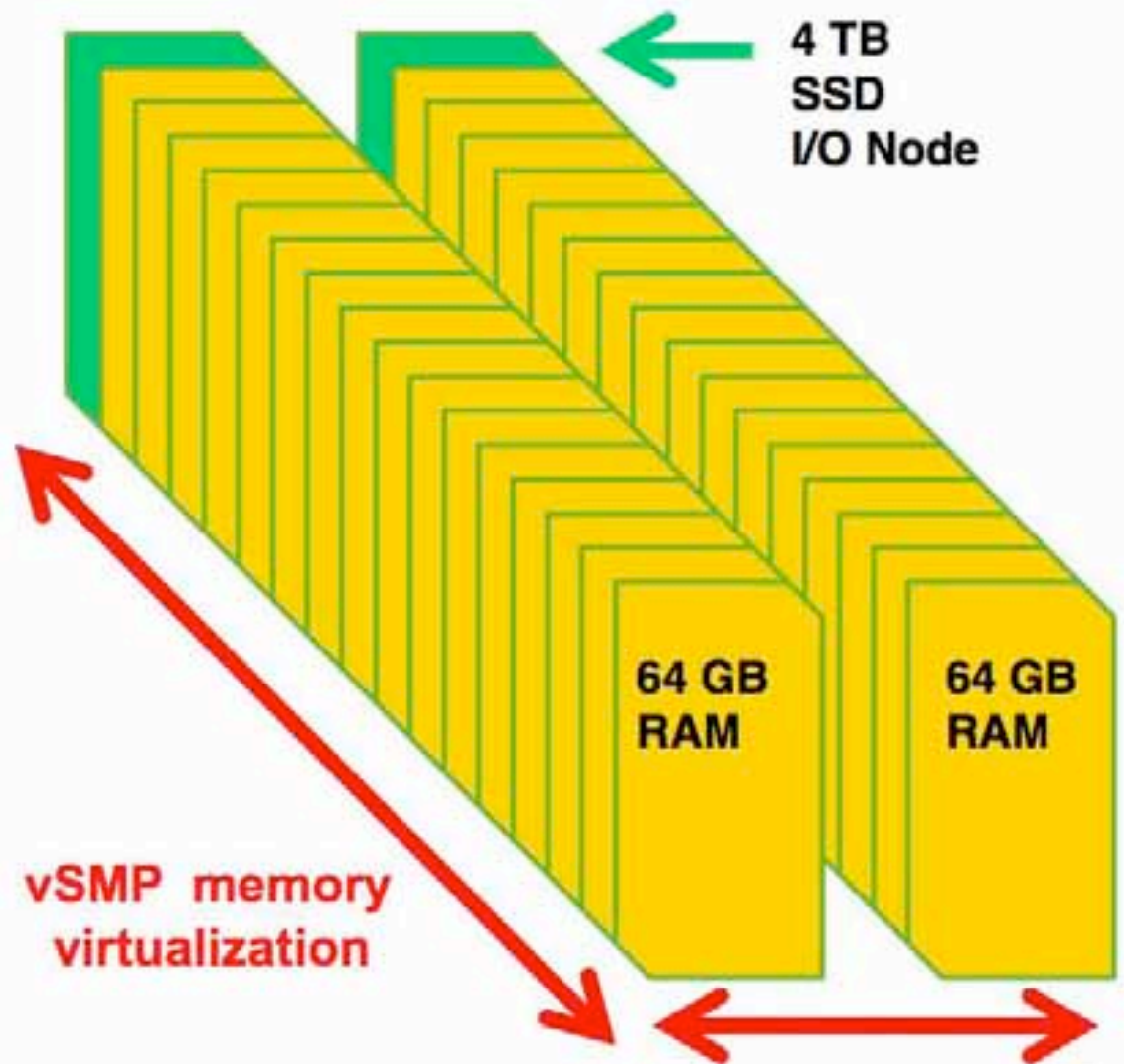
- Dual processor Intel Sandy Bridge
  - 64 GB

## 2 Appro Extreme-X IO nodes

- Intel SSD drives
  - 4 TB ea.
  - 560,000 IOPS

## ScaleMP vSMP virtual shared memory

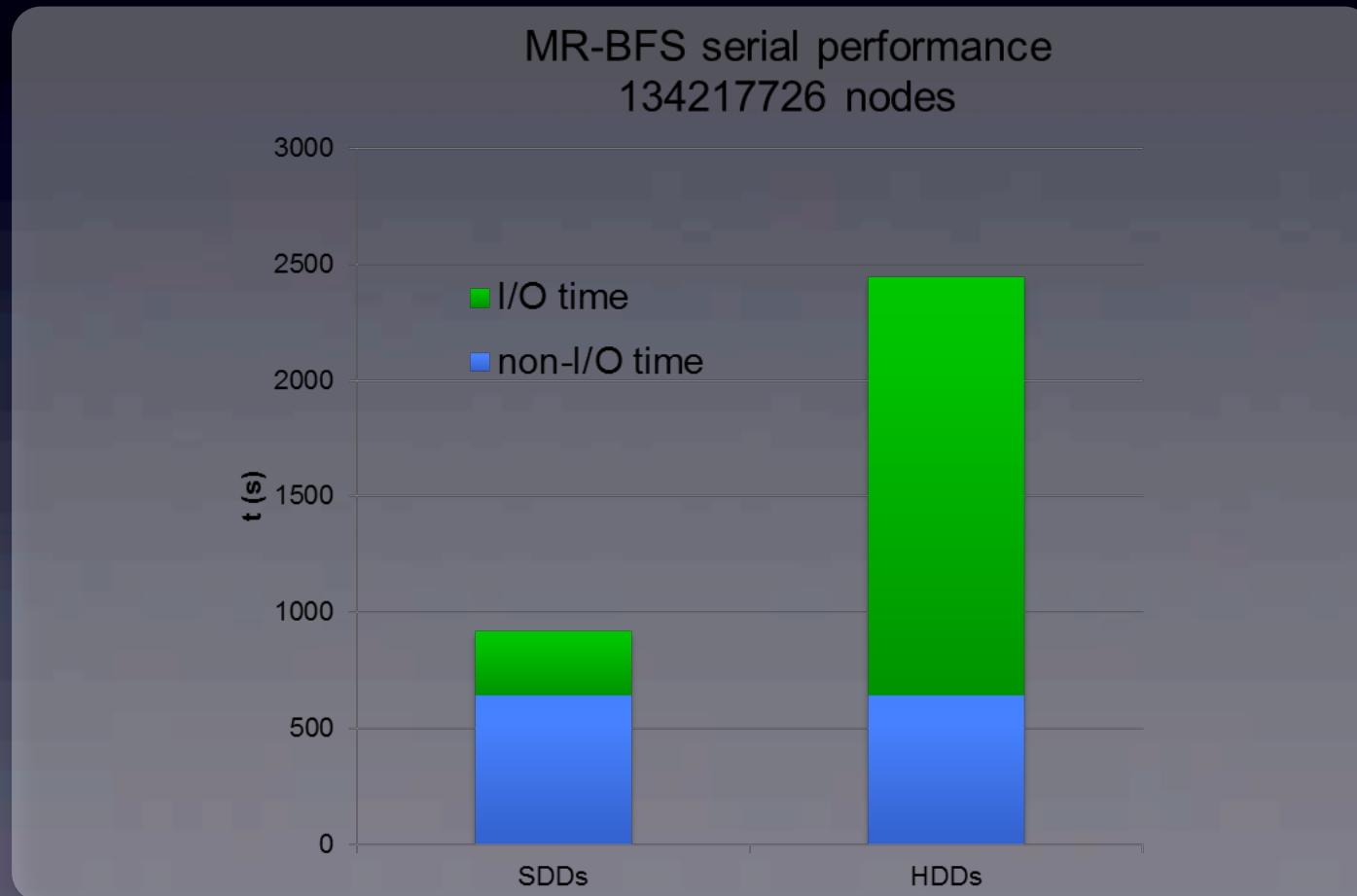
- 2 TB RAM aggregate
- 8 TB SSD aggregate



Full machine is 32 supernodes interconnected by dual-rail QDR IB in 3D torus.



# Gordon BFS Performance



*6.5x Improvement - Available now through XSEDE ([www.xsede.org](http://www.xsede.org)).*



# Calxeda/HP Moonshot



*“The EnergyCore is a single chip with a Cortex-A9 ARM processor running between 1.1GHz and 1.4GHz. The chip includes 4MB of cache, an 80-Gigabit fabric switch and a management engine for power optimization. Servers with the chip, 4GB of memory and a large-capacity solid-state drive [SATA] draw 5 watts of power. Besides using a low-power ARM processor, Calxeda has cut down chip power consumption by integrating key server components.”*

- Agam Shah, Computerworld, Nov 11, 2011



*Cyberbricks and Gordon are real*

*(and 6-10x is great but still constrained by I/O architecture)*

*What if we do “make believe” computer architecture?*

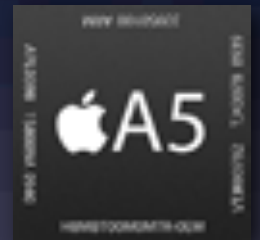


# Power Miser Devices: Apple A5 to CI

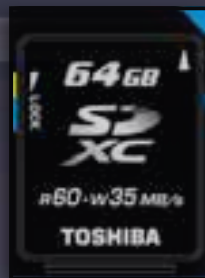
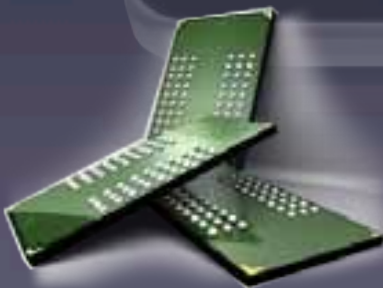
- A5



- ▶ SoC/PoP
- ▶ ARM/GPU/USB 2.0/Flash cntrl.
- ▶ 512 MB
- ▶ 10 Gflops? at 1W?
- ▶ 64 GB NAND Flash



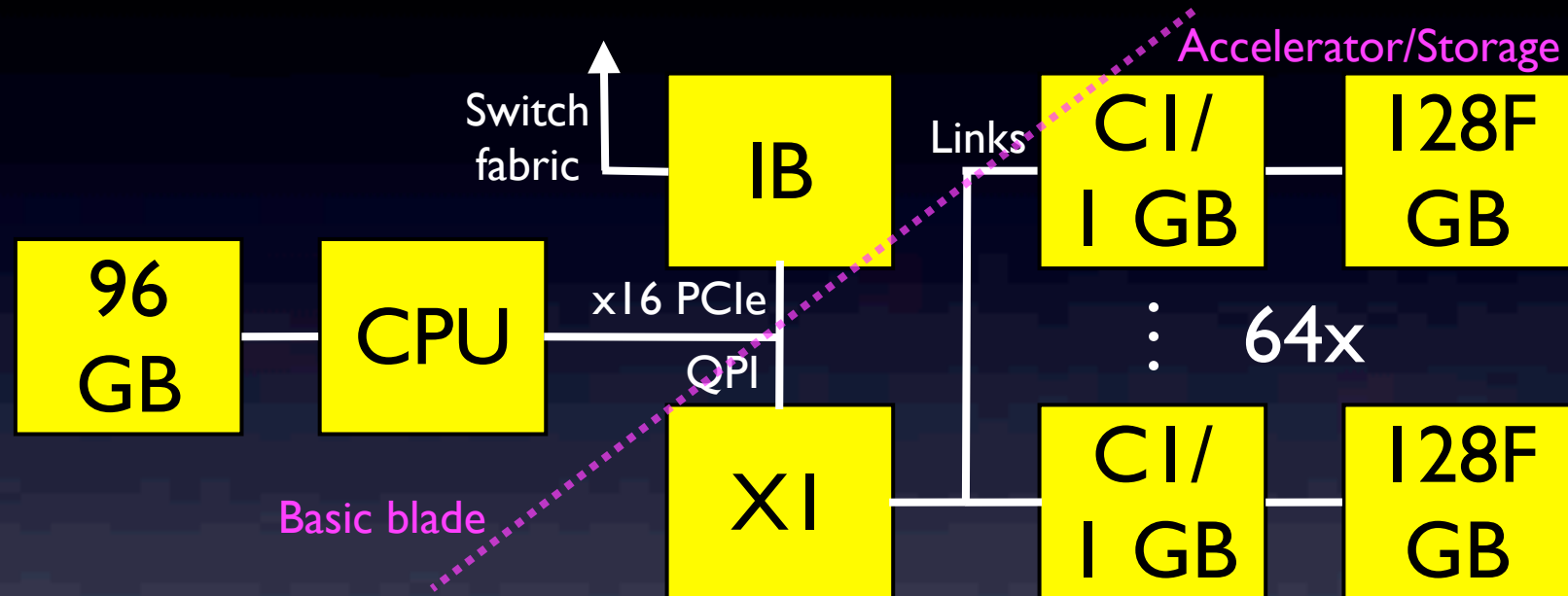
- CI



- ▶ Proc./Accel./Link/Flash engine
- ▶ 1 GB LDDR2 (64 bit wide) PoP
- ▶ 64 Gflops at 6 W + 2 W (1 Ghz)
- ▶ 128 GB NAND Flash RAID
- ▶ All existing IP; need < 1 year



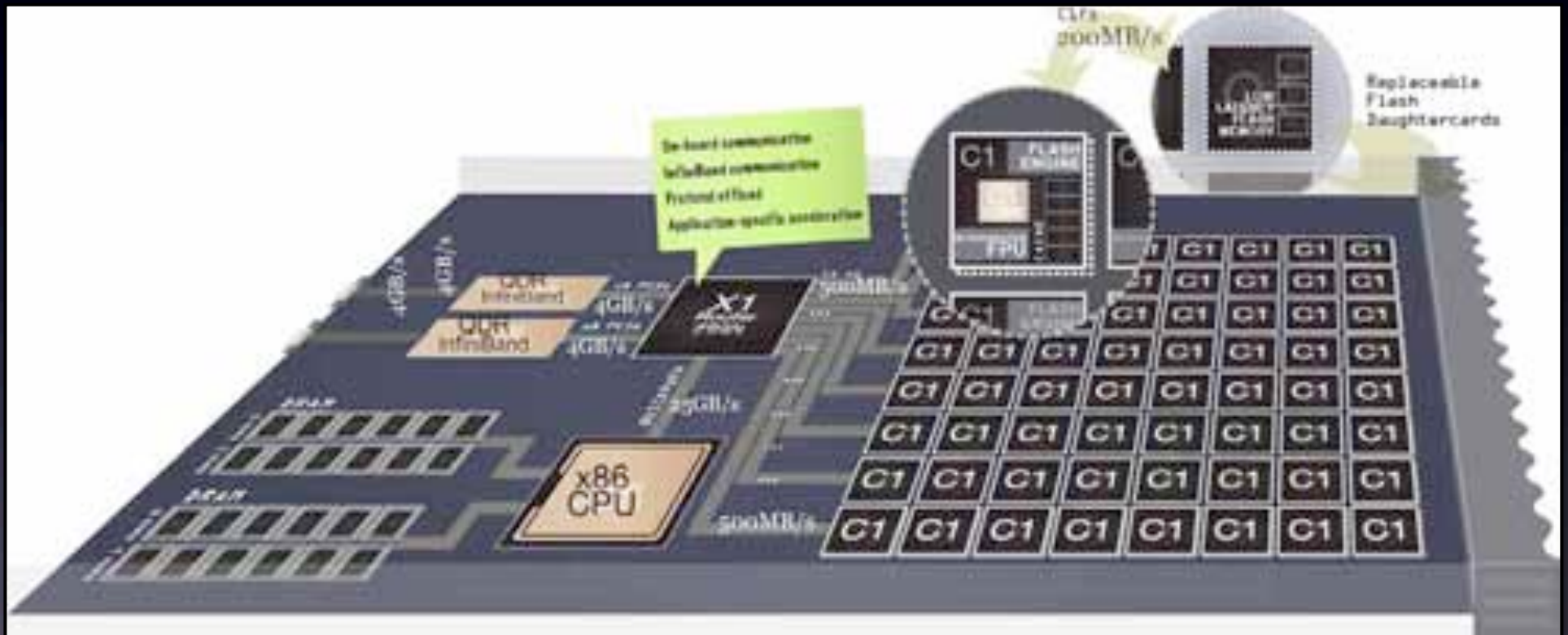
# A More Extreme Approach: FlashBlades



- “XI” is an FPGA switch for CI array & QPI to CPU & PCIe to IB
- The CPU orchestrates *abstractions*; to the CPU the array looks like:
  - ▶ A ~6 TB, 25 GB/s (burst), 50 us, || disk (file system, triple stores)
  - ▶ A ~4 TFlops accelerator (OpenCL with embedded triple stores)
- This all fits on a standard blade (2 sides) and uses commodity IP
  - ▶ Draws about 600 W and is 100x faster on disk operations



# FlashBlade Packaging



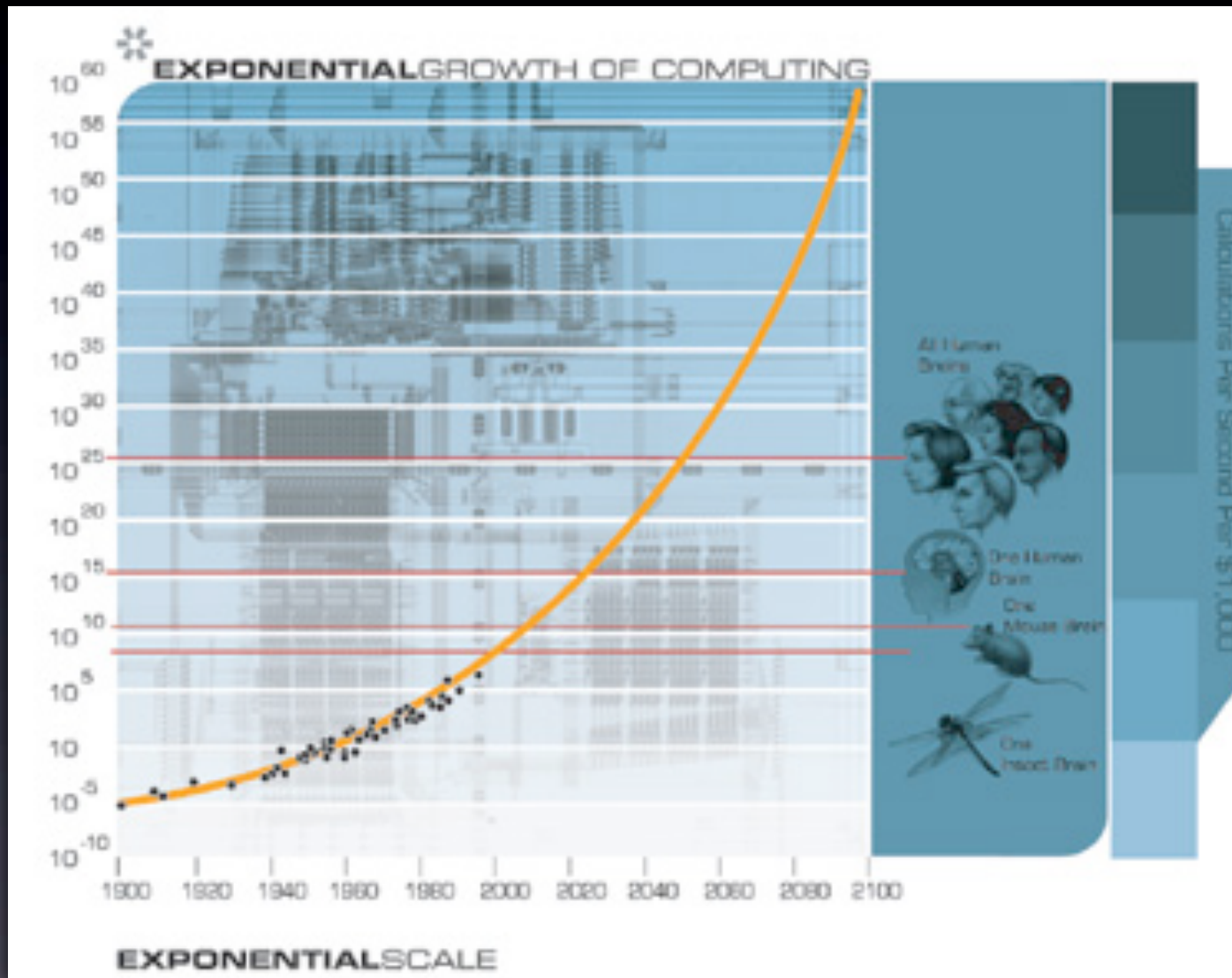
Stalzer, Workshop on Architectures and Systems for Big Data (ASBD), June 2012



## *Another Big Data Implementation Technology*



# It Only Takes $10^{16}$ Ops (?)



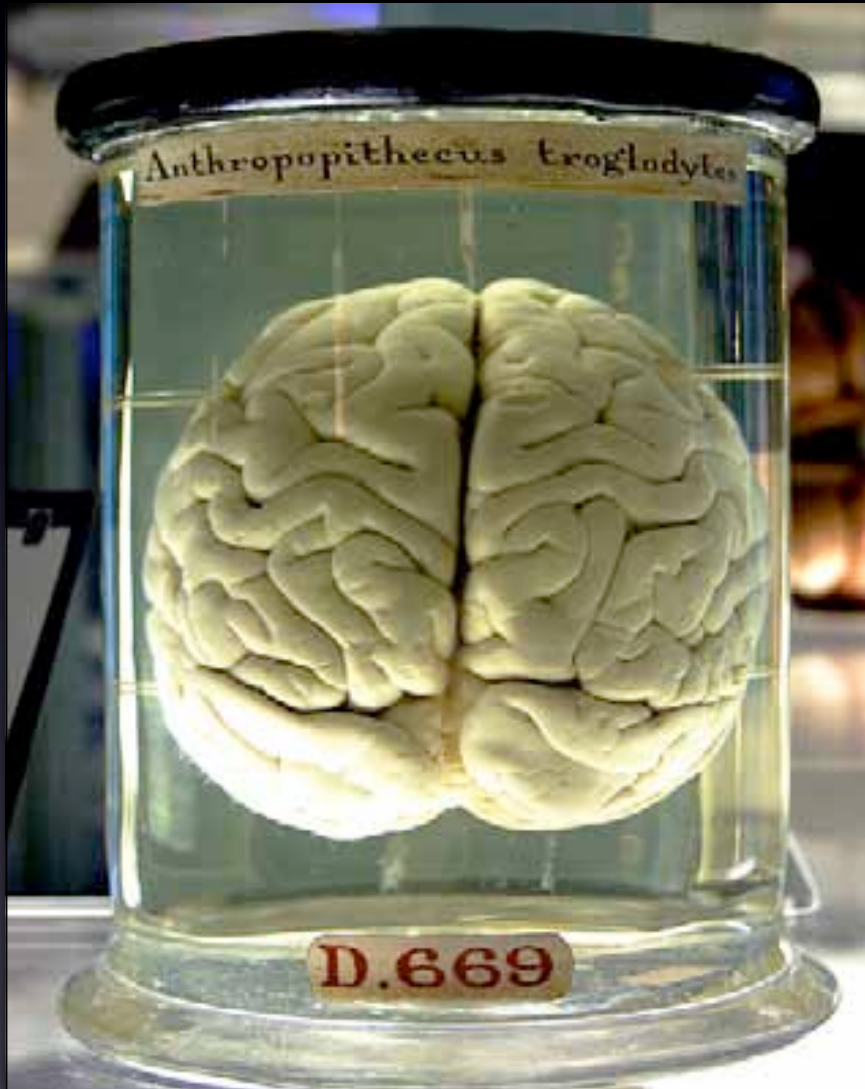


# A $10^{16}$ Flops Engine

- Quantity: Need about 200 servers (14 blades at 7U each)
  - ▶ Big IB switch fabric too (168 IB ports/rack)
  - ▶ Volume (need lots of air): 2,000 ft<sup>3</sup>
- Flash memory: ~17 PB
  - ▶ SDSC runs 18 PB of tape storage (1,000's of scientific data sets)
  - ▶ Constraint: no more than 1,000-2,000 writes/chip/day?
  - ▶ Data ingestion and reflective analysis by engine
  - ▶ Checkpoints in <10 s
- About 10 Pflops at ~2 MW (does not include cooling)
- Cost unknown due to rapidly dropping flash costs, new packaging, and other economies of scale



# Comparison to a Natural Big Data Engine



- Operations: 10 Pops (1x)
- Memory: 1 PB (0.06x & forgets)
- Bandwidth: 1 PB/s? (0.5x - 12x - 30x)
- Packaging: 0.25 ft<sup>3</sup> (8,000x)
- Power: 25 W (80,000x)!
- *Where's the algorithm?*



# Some Clues from Biological Systems: Packaging

- VLSI has a few interconnect layers (many more process layers)
- Fractal topology:  $\log(\#nodes): \log(\#ext. \text{ edges})$  scaling with box size
- Data from Bassett et al., PLoS Computational Biology, Apr 2010:

Network	$D_{\text{Euclidean}}$	$D_{\text{Fractal}}$
VLSI	2	3.81+-0.64
C. elegans	3	4.42+-1.53
Brain (MRI)	3	4.12+-1.55



## *Concluding Remarks*



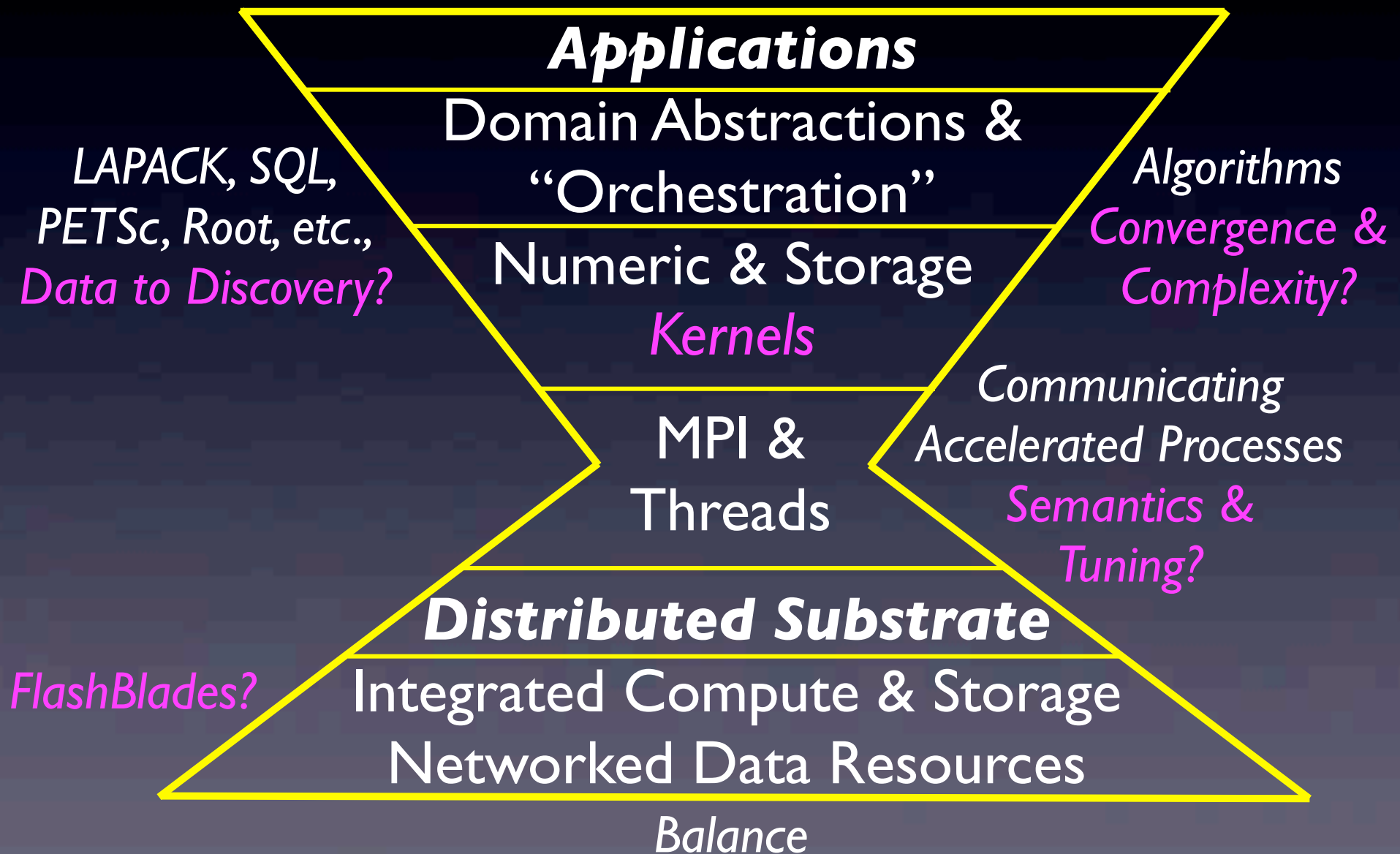
# Socket Archipelago (2018)

	Cluster	Socket
Parallelism	100,000	10,000+
Rel. Latency	~1,000	1

- Parallelism is becoming dramatically bimodal
  - ▶ MPI (or process) is essentially island level parallelism
  - ▶ Threads are tribe level parallelism and much much faster
  - ▶ What if all threads want to talk with another island at once?
- Must have very large L2 cache on socket
- Non-volatile storage must be integrated too (stacking)
- *Analogy: cortical columns*

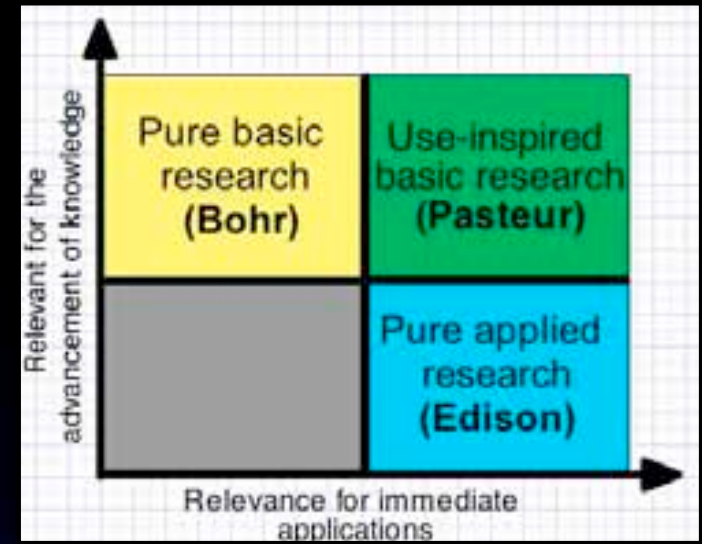


# Engine Software Architecture





# Pasteur's Quadrant



- Computer Scientists:
  - ▶ *Develop better abstractions to ease massively parallel programming*
  - ▶ *Improved algorithms*
  - ▶ *Manage asynchrony and (un)reliability*
- Material Scientists:
  - ▶ *Better insulators*
  - ▶ *Higher dimensional interconnects*
  - ▶ *New solid state storage technologies*
  - ▶ *What's the next S-curve (graphene?)*
- *Both: Think in terms of what can be done with a shrinking  $10^{16}$  ops system (socket archipelago)*



# Working Group

## (Hardware Trends in Computing)

- Computational requirements for scattering science
  - ▶ Data Volume, Velocity and Variety
  - ▶ Simulation workload
  - ▶ 5 yr trends and algorithmic scaling
- Cyberinfrastructure
  - ▶ Present capabilities
  - ▶ 5 yr trends
- Gaps?
  - ▶ General purpose or dedicated resources?