Current & Planned 2012 Lattice QCD Facilities @ Jefferson Lab

Chip Watson

Scientific Computing Group

Quick Outline

- Hardware Overview & Recent Changes
- Operations Report
- 2012 Conventional Infiniband x86 Cluster
- 2012 Accelerated Cluster Plans

Hardware Overview – IB Clusters

Infiniband Clusters

- "9q" 320 nodes dual Nehalem (@ 1.96 Jpsi)
- "10q" 224 nodes dual Westmere (@ 2.0 Jpsi)
- Configured as 1 set of 1024 cores, 13 sets (racks) of 256 cores
- All nodes have QDR Infiniband; 256 core sets have full bandwidth, large set has 2:1 switch oversubscription
- Dual QDR uplink to the file system

One of these 17 racks contains GTX-285 GPUs, and is dual use with the GPU cluster.

Hardware Overview – GPU

GPU Nodes

- 118 quad GPU, dual Nehalem/Westmere, 48 GB memory

GPU Configuration	*Infiniband Configuration*
36 quad C2050/M2050 (ECC)	8 @ dual rail QDR, 28 @ ½ QDR
32 quad GTX-580 **new!**	½ SDR
40 quad GTX-480	½ SDR
10 quad GTX-285 (weight 0.4)	½ SDR

- 34 single GTX-285, dual Westmere, 24 GB memory, full QDR
  (shared with Infiniband cluster (1 rack of 10q), with GPU having priority)

Users may select to have ECC memory, or 50% higher single precision performance, or 4x CPU cores + 2x memory per GPU. All of these options have identical weight. Only the quad GTX-285 has lower weight due to lower performance and no offsetting advantages.

Hardware Overview – Disk

4 name spaces

/home (small, user managed, *on older Dell system, soon to be upgraded*)

/work (medium, user managed, *on Sun ZFS systems, soon to be upgraded*)

/cache (large, write-through to tape, auto-delete when 90% full, on Lustre)

/volatile (large, auto-delete when 90% full, on Lustre)

Lustre

- fault tolerant metadata server (dual head, auto-failover)
- 23 Object Storage Servers (OSS), all on Infiniband, > 4GB/s aggregate b/w
- 380 TB (usable) allocated to sum of /cache and /work
- will be expanded by 120+TB this summer for new allocations

Custom management software

- separate project quotas for /cache and /volatile
- sum of quotas exceeds capacity (any active project can exceed quota)
- triggers deletion when /cache or /volatile reaches target size (90% full); deletes files from groups over quota first, then proportional to quota

Operations

Summer 2011 Cyber Security Incident ☹

My Apologies!!!

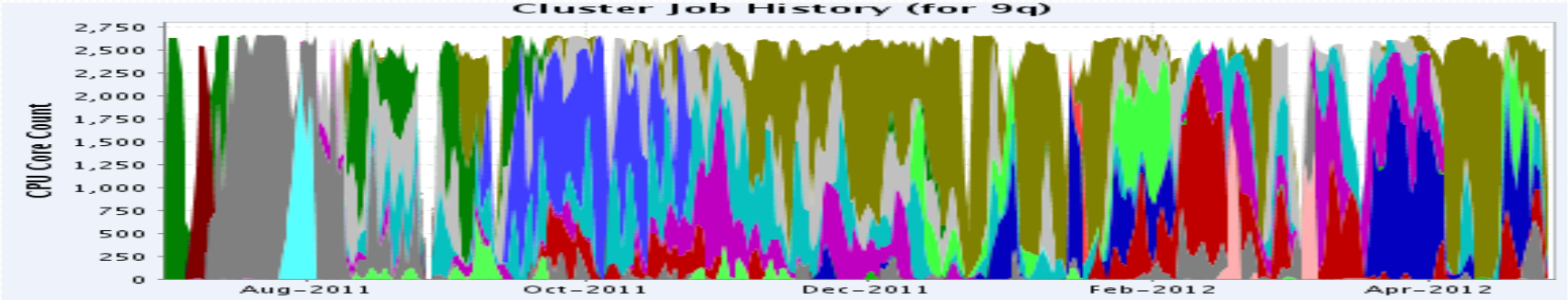
When the intrusion was detected, Jefferson Lab closed itself off from the internet except for email (no web). Later, white-listed hosts could connect via ssh. This happened at the worst possible time – just as we were transitioning to a new allocation year. To add insult to injury, one of our sys-admins left with 2 weeks notice for a higher paying position. It was 2 months before we were at anything resembling "normal". Fortunately, on-site users and a handful of users with early white-listed home machines were able to keep the USQCD computers busy and consume their allocations, otherwise cycles would have been lost.

Fair share: (same as last year)

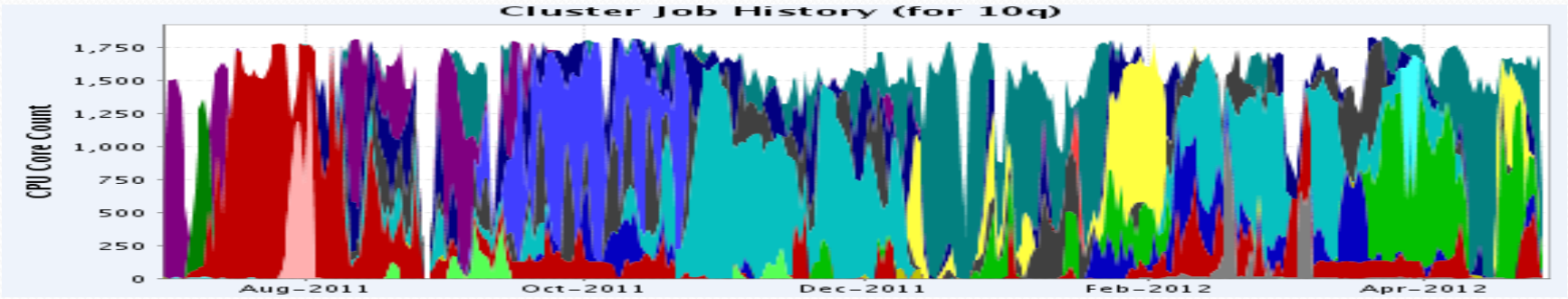
- Usage is controlled via Maui, "fair share" based on allocations
- Fair share adjusted every month or two, based upon remaining unused allocation (so those who quickly consumed their allocations later ran at zero priority)
- Separate projects are used for the GPUs, treating 1 GPU as the unit of scheduling, but still with node exclusive jobs

Infiniband Cluster Utilization

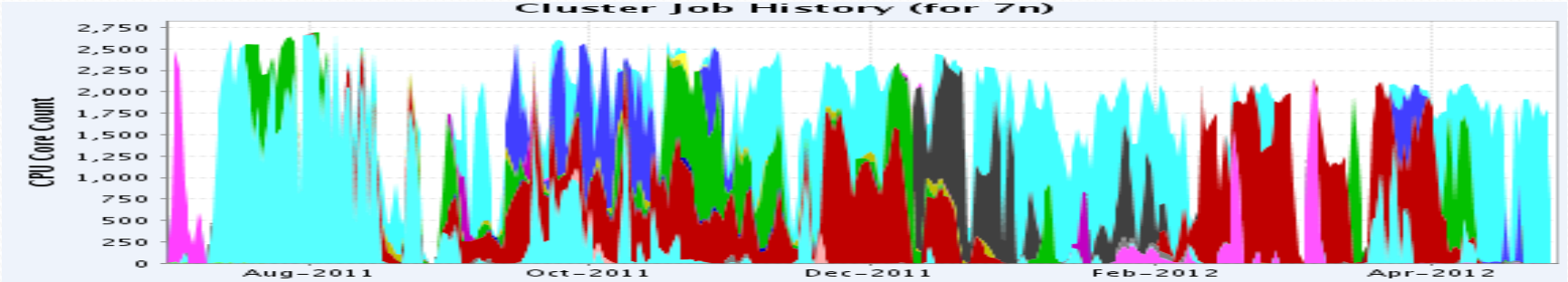
9q



10q



7n

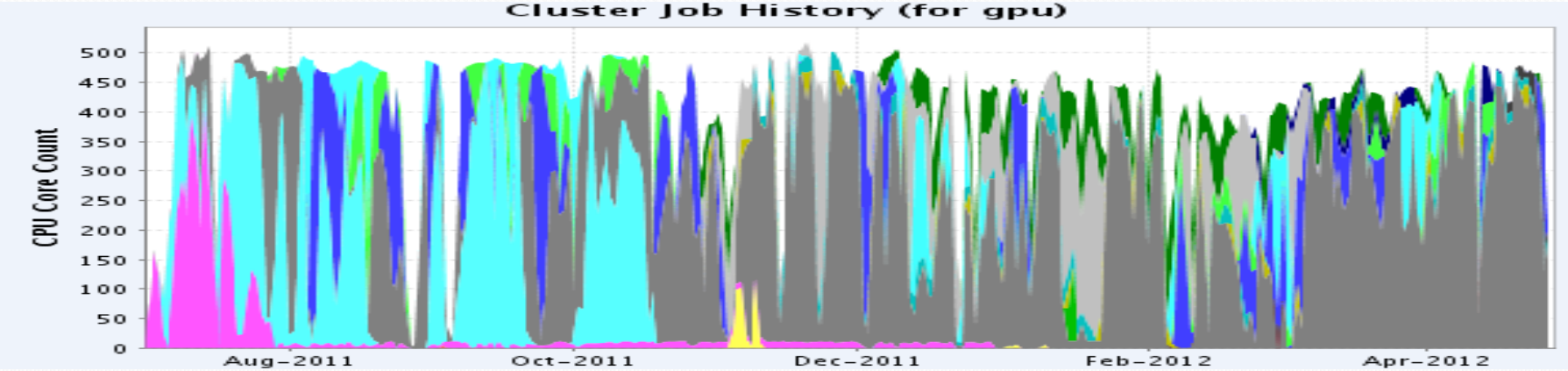


Colors represent users, but are not correlated between graphs.

2nd graph has fluctuations of 256 cores as 17th rack flips to/from GPU use.

Least popular 7n often underutilized (and will be turned off May 14).

GPU Utilization (Un-normalized)



- Occasional dips in utilization, but generally heavily used
- The sag in February 2012 was for debugging an upgrade from GTX-285 to -580, which yielded > 10% additional capacity

  Although only half of the 40 upgraded systems went quickly into production, this was still a capacity increase as each was 2.5x faster; eventually 30 went into production, and the other 10 were downgraded back to -285 and put into production, hence the return rise in March/ April for GPUs in use.
- Current effective performance: 74 Tflops (weighted by allocations)

Infiniband Cluster Usage – 105% of pace

Project Name	Allocation*	Hours Used	Annual Pace	Monthly Pace	Hour Remaining	Overused
NPLQCD	24,300,000	15,234,410	76%	94%	9,065,590	0
Spectrum	15,020,000	16,740,417	136%	70%	0	1,720,417
LSD	12,080,000	12,804,600	129%	80%	0	724,600
thermo	11,230,000	11,865,590	128%	56%	0	635,590
emc	4,430,000	2,688,501	74%	324%	1,741,499	0
Tcolor	3,750,000	15,424	1%	2%	3,734,576	0
ffss	1,980,000	759,481	47%	6%	1,220,519	0
TMD	1,950,000	3,969,897	248%	1054%	0	2,019,897
EigenSpect	900,001	1,057,687	143%	172%	0	157,686
EMspect	600,001	566,620	115%	0%	33,381	0
strangeness	60,001	38,736	78%	0%	21,265	0
NcSU3	20,001	10	0%	0%	19,991	0
jplattice	20,001	0	0%	0%	20,001	0
Total	76,340,005	65,741,372	105%	113%	15,856,823	5,258,190

Projects with allocations ending in “1” are Class C.
Lab is ahead of pace mostly because of low requests for Class C allocation.

GPU Cluster Usage – 112% of pace

Project Name	Allocation	GPU Hours	Annual Pace	Monthly Pace	Hour Remaining	Overused
Spectrumg	1,167,000	2,029,889	212%	299%	0	862,889
NPLQCDg	628,000	154,417	30%	0%	473,583	0
thermog	564,000	516,200	111%	92%	47,800	0
Tcolorg	436,000	135,225	38%	0%	300,775	0
emcg	416,000	141,274	41%	0%	274,726	0
lattsusyg	150,001	120,510	98%	153%	29,491	0
staggwme	100,000	165,270	201%	61%	0	65,270
gwuQCD	90,000	68,593	93%	169%	21,407	0
discogpu	55,000	21,259	47%	164%	33,741	0
charming	20,001	549	3%	0%	19,452	0
Total	3,626,002	3,353,187	112%	125%	1,200,974	928,159

Only 5% given to Class C; this plus 285 => 580 upgrade yielded high % of pace.
75% of projects are on track to consume their allocations.
Only 2 of the top 5 projects were able to use more than half of their allocations.

http://lqcd.jlab.org/, Project Usage 11-12

New: 2012 Infiniband Cluster

Reminder: the project decided to spend between 40% and 60% of the hardware funds on an unaccelerated Infiniband cluster, and the rest on an accelerated cluster, with NVIDIA Kepler as the reference target device.

In March JLab placed an order for 212 nodes (42%):

Cluster Name: **12s** == 20**12 S**andy Bridge (latest Xeon CPU)

- dual 8 core CPU 2.0 GHz; 1 core ~ 1.8 Jpsi cores
- 32 GB memory (dual socket, 4 channel, 4GB)
- Full bi-sectional bandwidth QDR Infiniband fabric (no oversubscription)
- Approx 50 Gflops/node, so ~10 Tflops (to be confirmed)

Delivery is expected late May for the first 6 racks. Early use in June (priority to unconsumed allocations). Production July 1. We are considering adding 2 additional racks (72 nodes).

USQCD Trends

- Applications that can exploit GPUs well have seen significant growth in performance over the last 3 years at modest cost to the project (22% of hardware budgets)
- Applications that need supercomputers are likely to see healthy growth in the coming year (ANL, ORNL, NCSA, …)
- Other applications are not seeing the same growth in performance

Each year, the LQCD computing project (s) must decide how to best optimize procurements for the community. The next step in this ongoing process is optimizing the use of the remaining 58% of 2012 funds.

Community Input

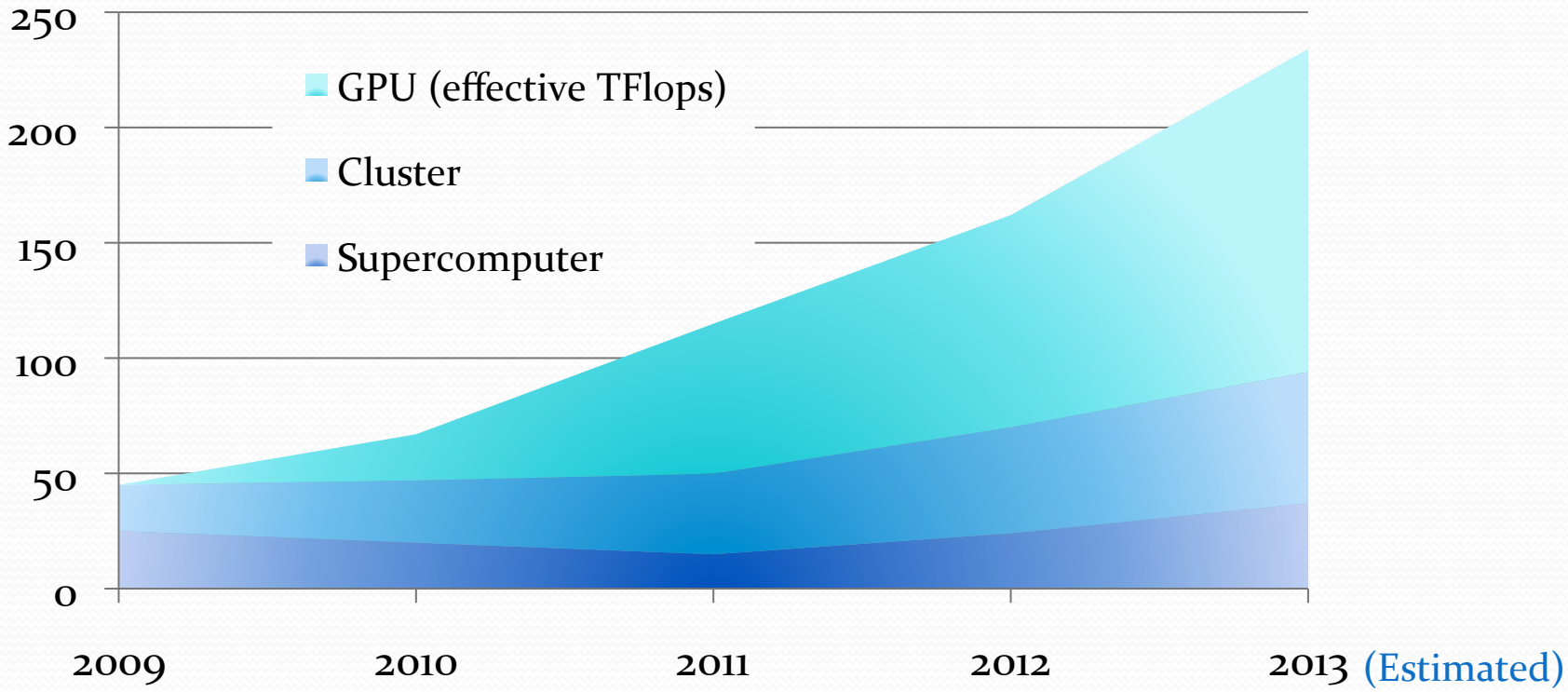
The project is guided by…

- Data obtained from the proposals
- Additional input from the Scientific Program Committee
- Input from the Executive Committee

and

Input from You!

USQCD Resources (effective TFlops)



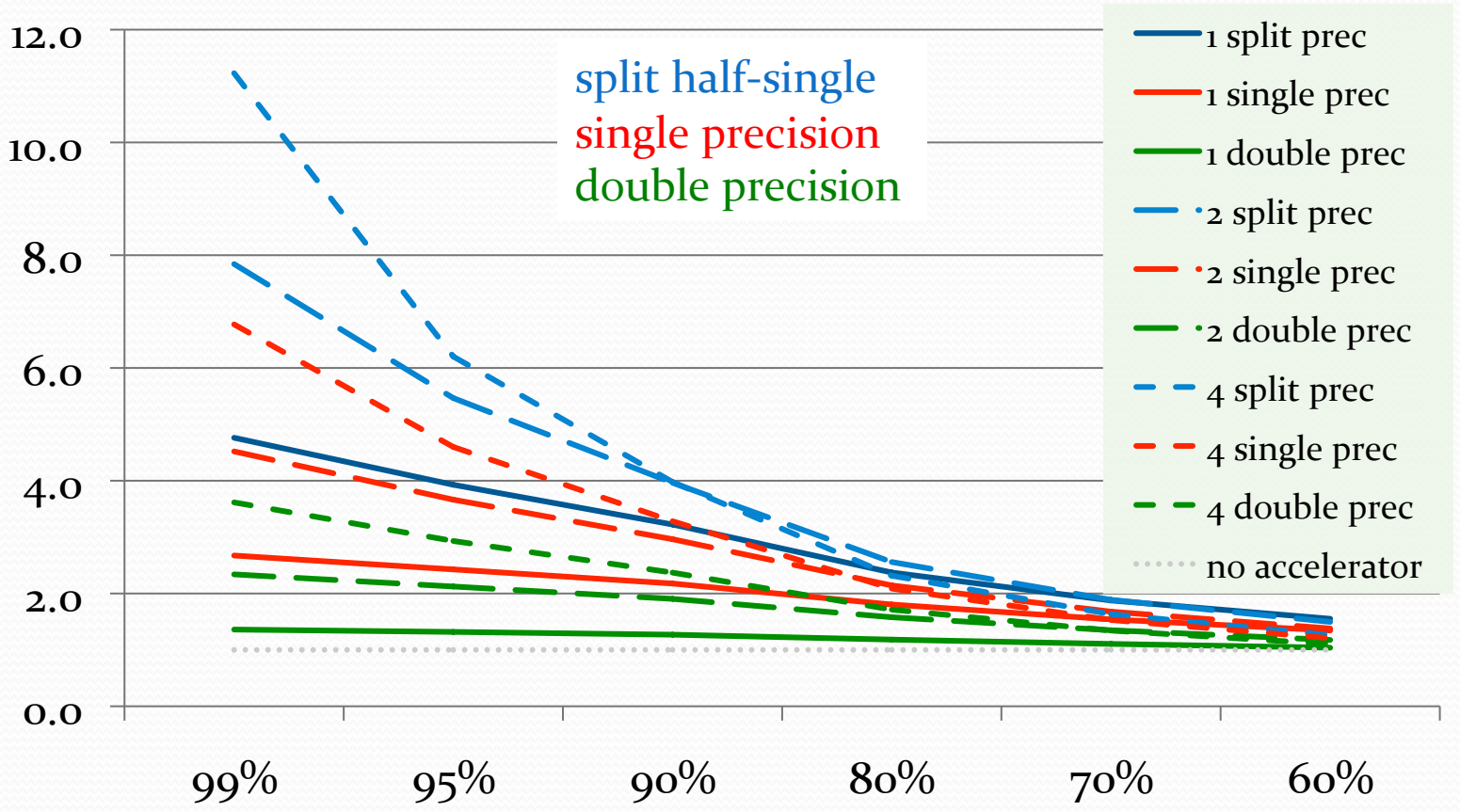
GPU Tflops is the equivalent cluster Tflops needed to do the same calculations.

Note: Supercomputer time does not include NSF, RIKEN, or other non-USQCD resources, which would probably double the displayed supercomputer time.

The GPUs have been a great success, providing more than half of the total flops for USQCD for the last two year.

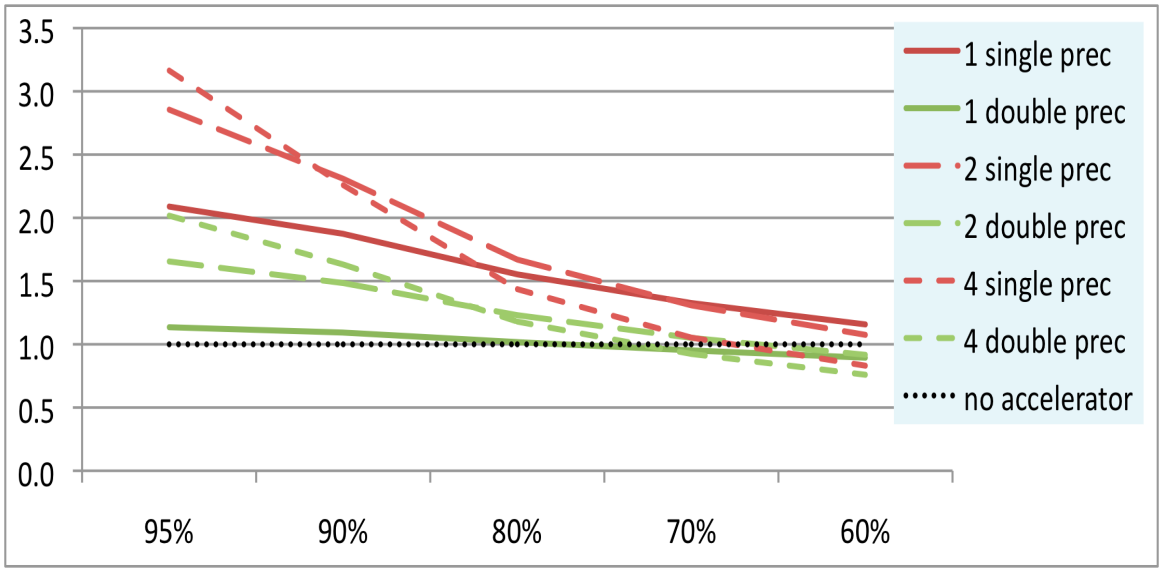
GPU Strengths & Limitations

Amdahl's Law and Tflops/$ Gain



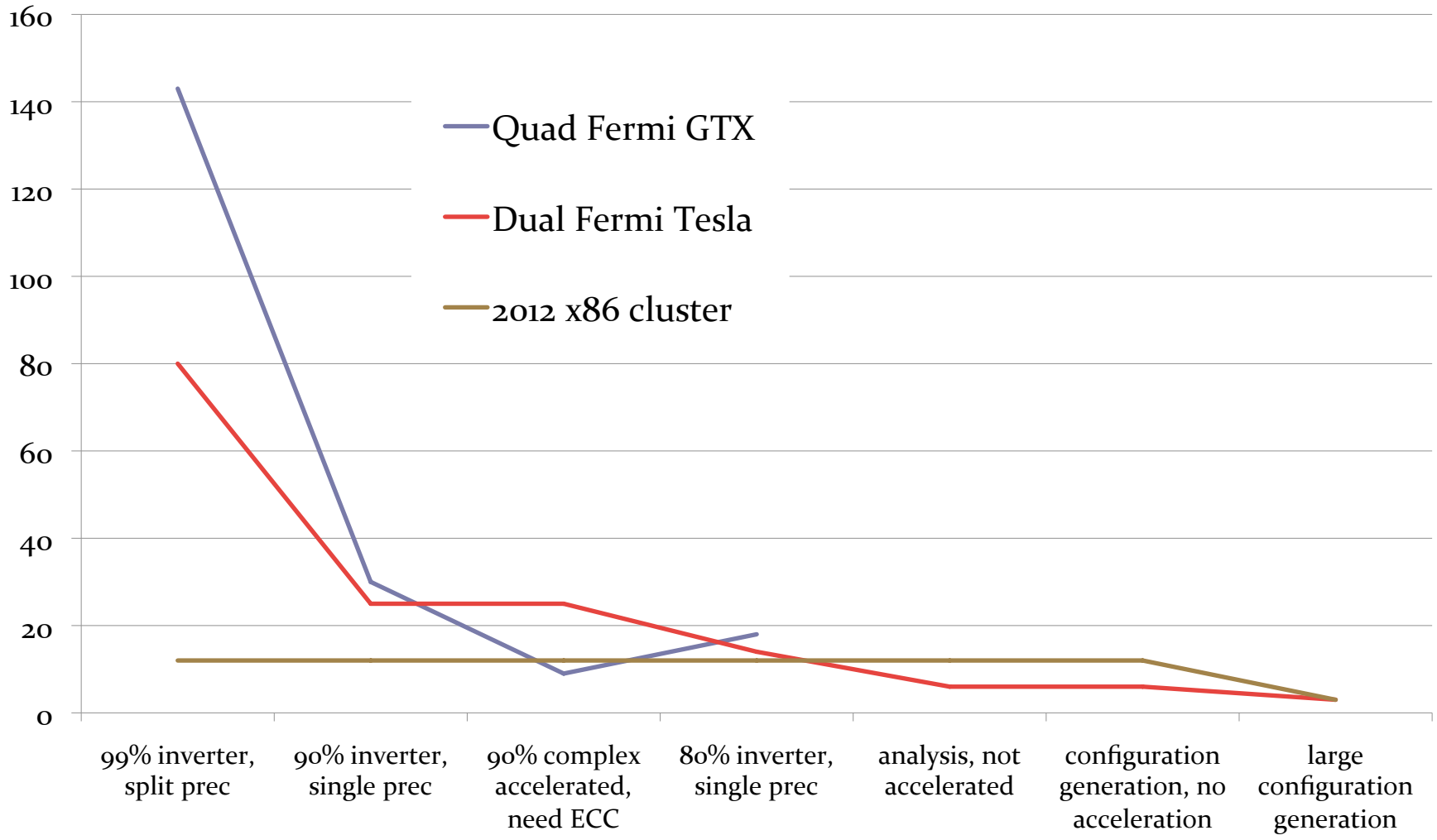
Accelerators work great when you accelerate > 90% of the code (e.g. inverters).
Gains shown are for inverters using GTX-580 with a quick test of correctness.

Amdahl's Law, for more expensive GPUs w/ ECC memory (smaller gains)



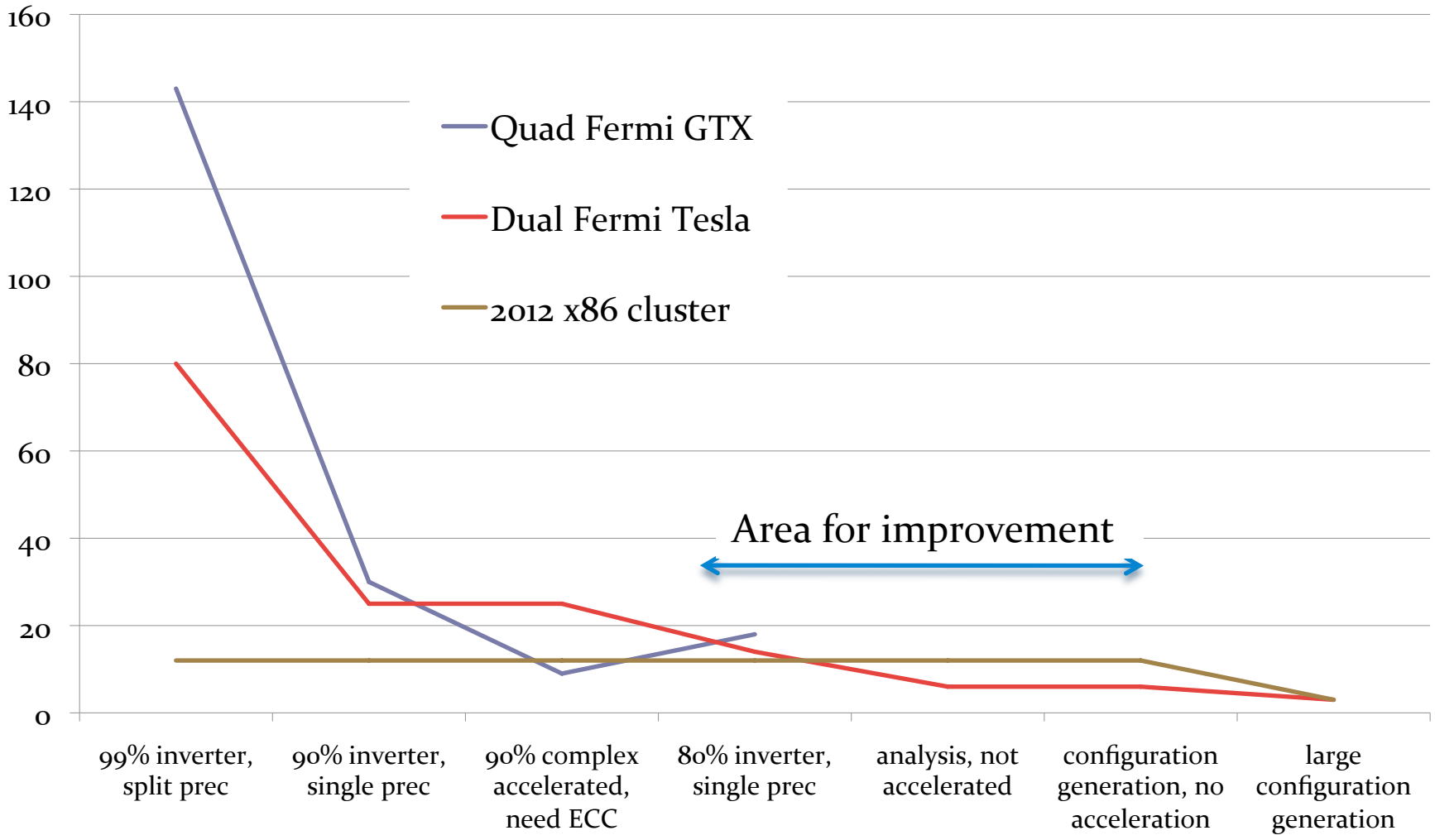
- For the more expensive Tesla GPUs, the requirement to accelerate almost all of the code is even more demanding. The 2x crossing point for single precision is around 85%, and for double precision it is around 95%.
- Data shown is for Fermi Tesla (C2050) at $1600/card vs. Sandy Bridge 2.0 GHz at $4000 per dual socket node (12s procurement).
- NVIDIA Kepler might do better, depending upon both performance and cost (tbd).

Price/Performance vs. Application



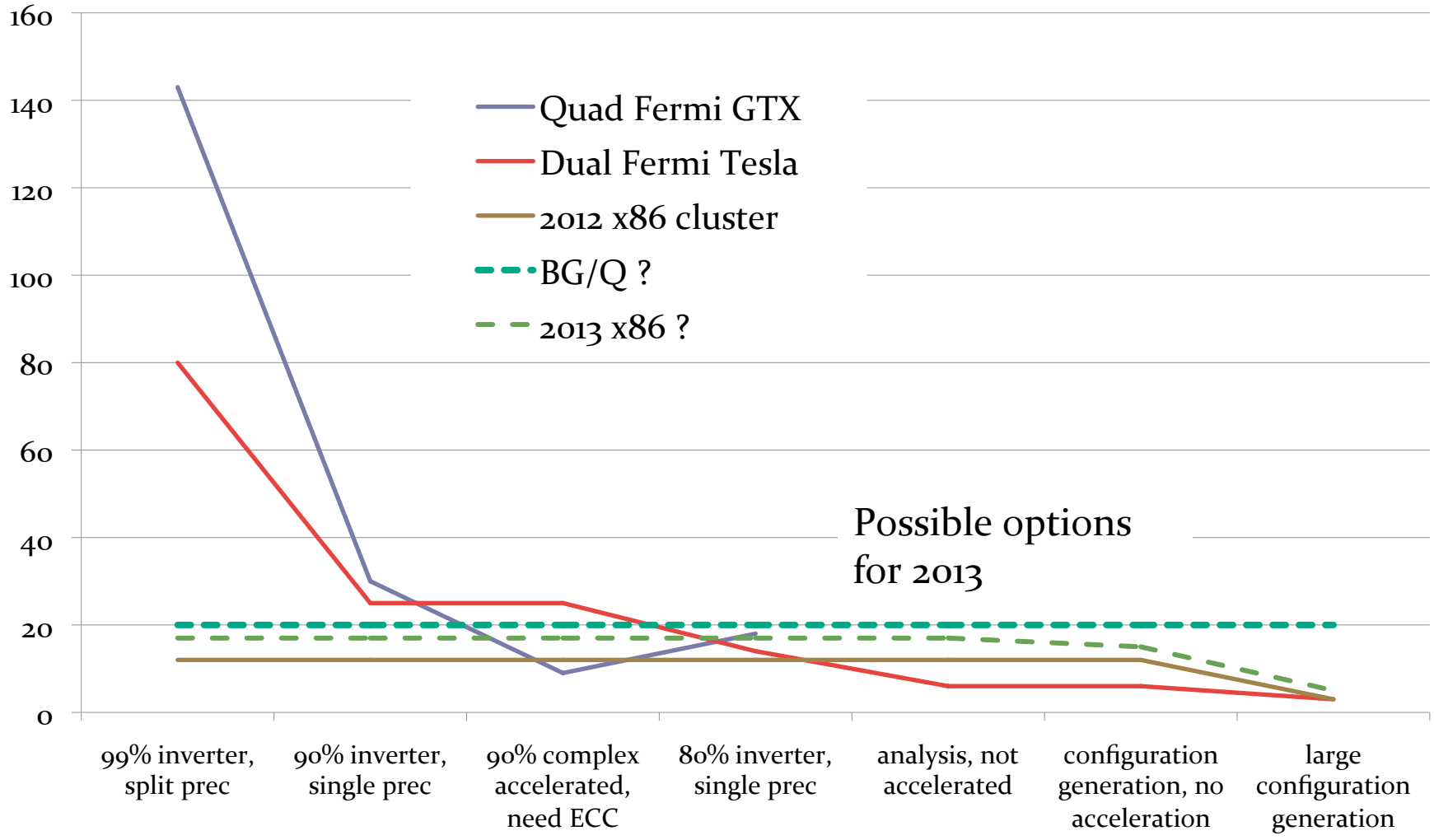
90% of the run time must be accelerated to make GPUs effective.

Price/Performance vs. Application



Spending 60% on conventional clusters will help in this range.

Price/Performance vs. Application

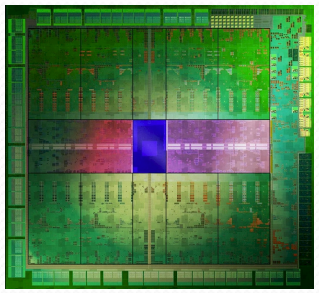


Moore's Law helps, raising the line 50% - 60% per year, but is slowing.

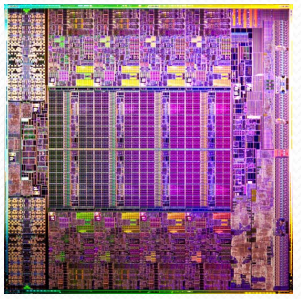
Multi-core Processor H/W Trends

(the following 4 slides courtesy of Balint Joo)

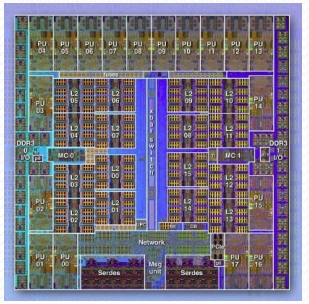
- More Cores: 16-64 cores per node
- Use of short vectors:
  - 4 SP / 2 DP (SSE)
  - 8 SP / 4 DP (AVX),
  - 4 DP (BG/Q-QPX)
- Hierarchical memory
  - L1 cache: small, low-latency, high-bandwidth
  - DRAM: high-latency, low-bandwidth
- Large Last Level Caches
  - Sandy Bridge: 20 GB Shared L3
- Non Uniform Memory Access (NUMA)
  - Between Sockets & Within Socket (AMD Interlagos)



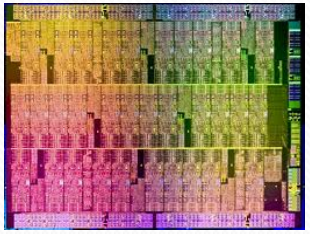
NVIDIA Kepler (1)
(The Register)



Xeon E5-2600
(legitreviews.com)



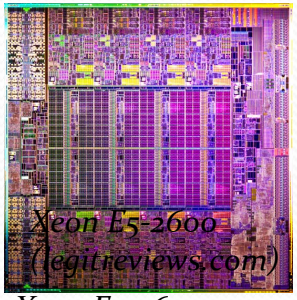
IBM BG/Q Die
(HPCWire)



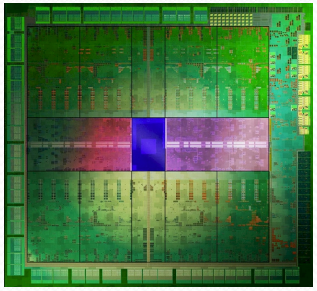
Intel MIC architecture
(techeta.com)

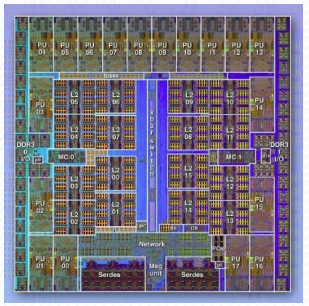
Multi-core Processor S/W Trends

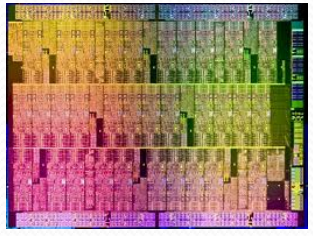
- More Cores: On-core threading (OpenMP, QMT, etc)
- Use of short vectors:
  - 'Vectorizable' C, #pragma hints
  - Compiler Intrinsics, Assembler, Code generators
  - 'Vector Friendly' data layout
- Hierarchical memory/BW constraints
  - Cache blocking,
  - Streaming Stores
  - Compression (e.g. SU3)
- Non Uniform Memory Access (NUMA)
  - threads must 'touch' data after allocation
  - Important to bind threads to cores carefully



Xeon E5-2600 (legitreviews.com)


NVIDIA Kepler (1) (The Register)


IBM BG/Q Die (HPCWire)


Intel MIC architecture (techeta.com)

Parallelization in Wilson Dslash

- Spins: SU(3) mat. x vec. for 2 spins at once (2-way)
- Directions: SU(3) mat. x vec. for 4 directions at once (4-way)
- Spins & Directions (8-way)
- For more than 8-way, we need to parallelize over sites

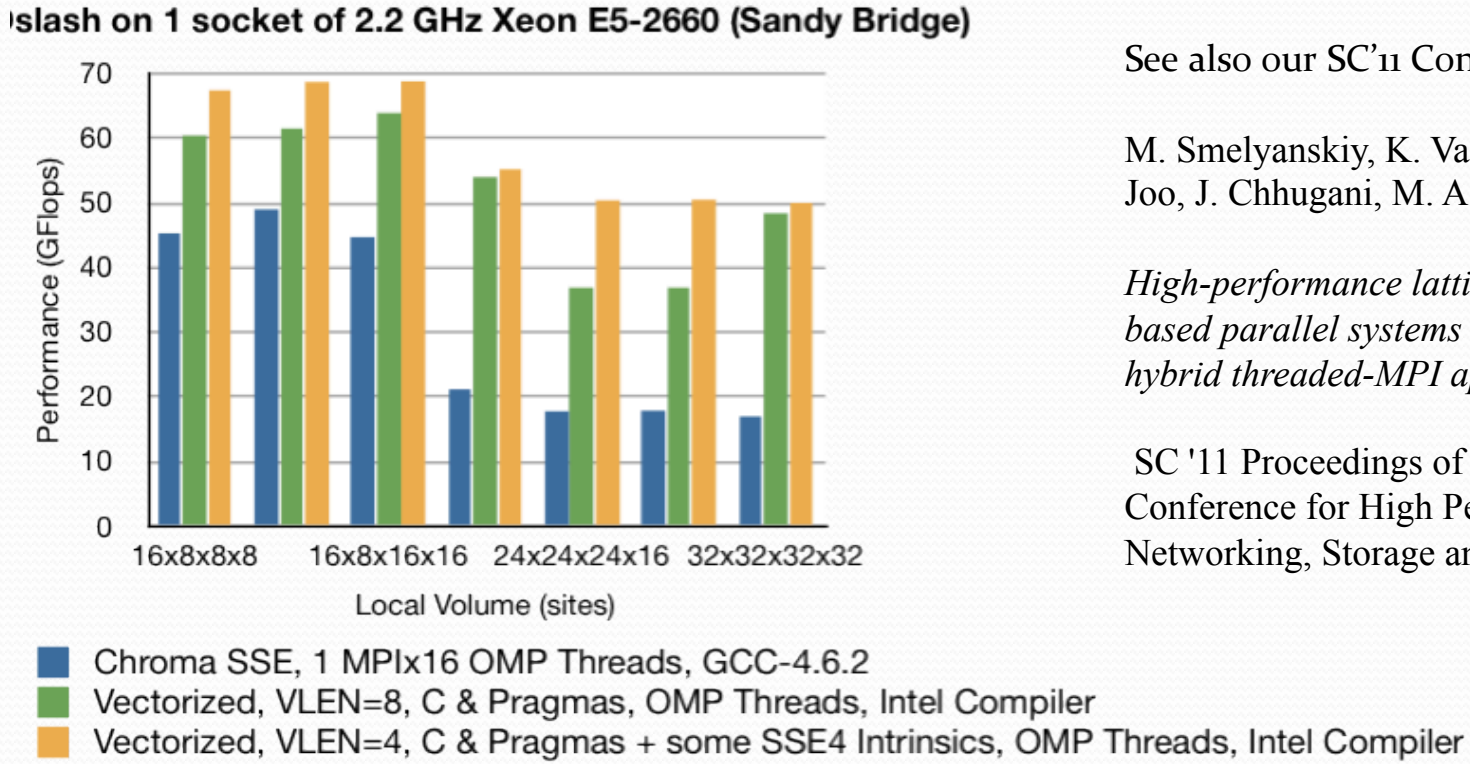
  => So called 'structure of arrays' (SOA) data layouts

```
// Natural layout: site-wise. E.g. Ns=4, Nc=3, NCmpx=2
float natural_layout[V_sites][ Ns ][ Nc ][ NCmpx ];

// QUDA layout (without padding):
// Split Nc x Ns into 6 x 4 floats,  4 x floats = float4
float4 quda_layout[6][V_sites][ NCmpx ];

// Blocked-Vector layout (without padding)
// Tune VECLEN: e.g. SSE=>4, AVX=8, Lx, Autotune
float vec_layout[V_sites/VECLEN][Ns][Nc][NCmpx][VECLEN];
```

Wilson Dslash on Sandy Bridge



See also our SC'11 Contribution:

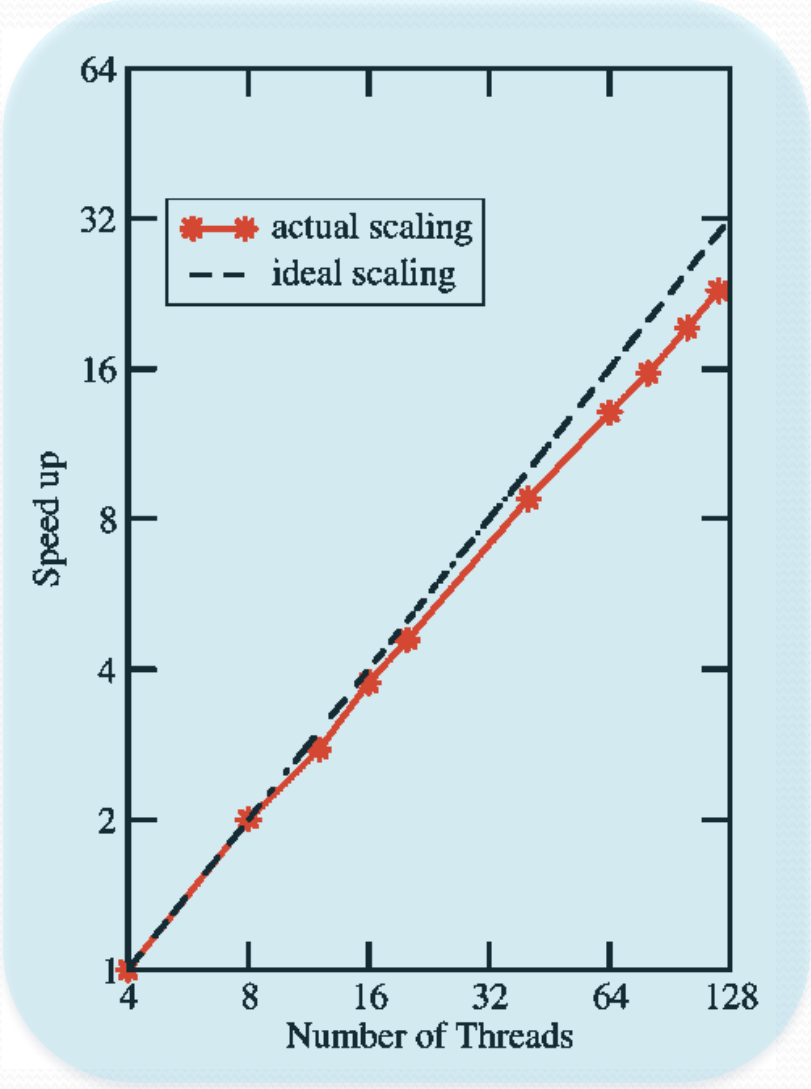
M. Smelyanskiy, K. Vaidyanathan, J. Choi, B. Joo, J. Chhugani, M. A. Clark, P. Dubey,

High-performance lattice QCD for multi-core based parallel systems using a cache-friendly hybrid threaded-MPI approach

SC '11 Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis

- Over 2x current Chroma performance for larger problems
- For VLEN=8, further optimizations possible with AVX intrinsics
- Collaboration with M. Smelyanskiy, Intel Parallel Computing Labs
- Expect similar benefits on most current CPUs (x86, AMD, BG/Q,...)

MIC – Many Intel Cores



MIC – pronounced ‘Mike’

- Many x86 cores, 512 bit wide vectors
- MIC will power the 10 Pflops NSF Stampede at TACC
- JLab is part of MIC Software Dev Program
  - Working on Highly Optimized Wilson Dslash, aiming for a High Performance Clover Solver (‘extreme programming’)
  - Also deployment of Chroma + Analysis software (‘regular code’)
- Chroma built & deployed in <1 day
  - tuning and optimization will take longer
- Collaboration with M. Smelyanskiy, Intel Parallel Computing Labs

MIC, Many Intel Cores

Jefferson Lab is a participant in Intel's MIC Software Development program, using Knights Ferry PCIe cards (they look like GPUs). KNF is a prototype of an upcoming MIC processor called Knights Corner to be deployed as part of the TACC 10 Pflops "Stampede" system.

The optimizations needed to achieve good performance for the dslash operator on x86 cores (Westmere, Sandy Bridge), a project that Balint has been doing in collaboration with Intel, are exactly the same optimizations needed to get good performance on MIC. Intel's tools report extensive data on success or failure to vectorize loops (very helpful).

Knights Ferry has "greater than or equal to" 32 cores, with 4-way hyper-threading, and a vector length of 512 bits (16 floats). It is an x86 processor on steroids for pure flops.

Knights Corner is the production version coming in <less than 1 year>

YACA? (Yet Another Computer Architecture)

- Is the potential worth pursuing?
- With growth in supercomputers, and with GPUs making inverters cheap, is it time to address the lagging middle?
- Will compilers take care of this, or do we really have to change our software?

Merging ARRA & LQCD-ext

LQCD ARRA and LQCD-ext have worked out the details to merge operations into the LQCD-ext project effective the beginning of FY2013 (a change request will be submitted).

As part of this step, the ARRA project will end at the end of this fiscal year.

Extrapolating labor costs through September, there remains approximately $150K for a final set of hardware enhancements, and discussions are now underway as to the best option for these funds.

A MIC testbed is being strongly considered.

MIC Testbed

The remaining ARRA funds could be used to procure an early testbed for MIC hardware. Users who are willing to work on optimizing their software for longer vectors could use this resource to good effect, enhancing USQCD's aggregate performance for that part of our application space not as well served by GPUs.

As a testbed, it would initially be free to users, with a bias towards those underserved by GPUs. Once its value is established, the MIC cards could be assigned a Jpsi core rating, but with charges divided by 2 so early users see a gain.

Procuring this testbed would be contingent upon proving that real applications could be ported to MIC with better than x86 price performance in under 6 weeks.

Time for your input:

- Do you have input on the x86 / GPU split for this year?
- For those of you with x86 allocations, would you be interested in investigating upgrading your allocation by trading in x86 core hours in exchange for MIC hours?
  - must have a large workload not currently addressed by GPUs
  - must be willing to invest one week in software development in the coming 2 months (working with Jlab staff)
  - must be open, if successful, to exchange 1M hours for a 2x performance gain on MIC nodes when/if they become available during this allocation year

Please grab me in the hall, or send an email!