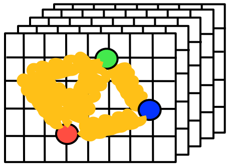
Enabling Lattice QCD Calculations using Graphics Processors

Bálint Joó, Scientific Computing

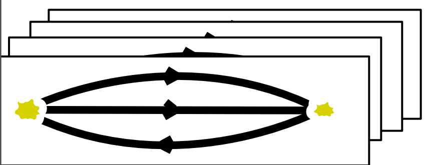
Science And Technology Review
JLab, May 9-11, 2012

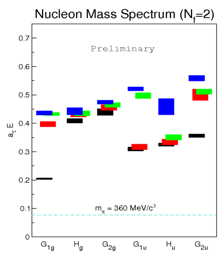
Large Scale LQCD Simulations

- Stage 1: Generate Configurations
  - configurations generated in *sequence*
  - *capability computing* needed for *large lattices* and *light quarks*
  - INCITE, collaborating institutions



- Stage 2a: Compute quark propagators
  - *task parallel* (per configuration)
  - capacity workload (but can also use capability h/w)
  - USQCD National Facility Clusters
- Stage 2b: Contract propagators into Correlation Functions
  - determines the physics we see
  - complicated multi-index tensor contractions

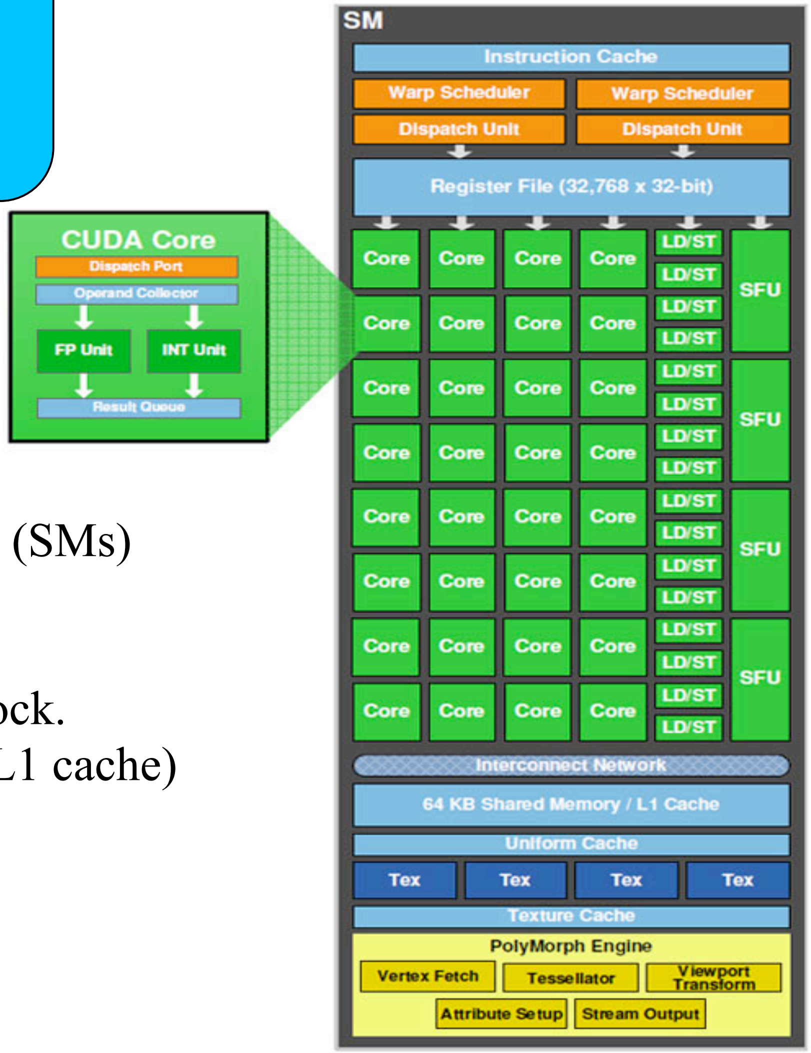




- Stage 3: Extract Physics
  - on workstations, small cluster partitions

Anatomy of a Fermi GPU

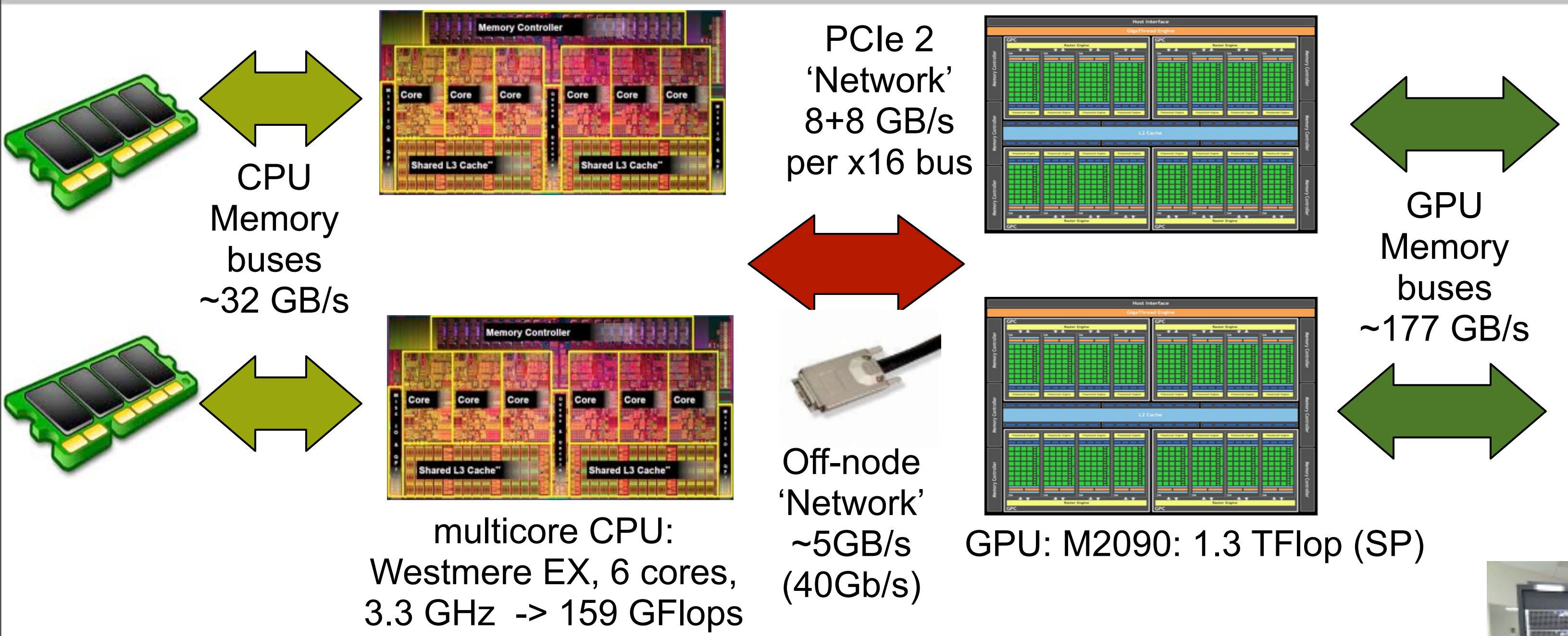
Tesla M2090;
512 CUDA cores
x 2 Flops/clock
x 1.3 GHz
= 1.33 Tflops (SP)



Streaming Multiprocessor (SM)

- NVIDIA GPU consists of Streaming Multiprocessors (SMs)
- SMs provide:
    - registers (32K 32-bit)
    - CUDA cores (32 per SM) – 1 SP mul-add per clock.
    - 64 KB Shared Memory (configured as memory/L1 cache)
    - Special Function units (for fast sin/cos/exp etc)
    - Hardware barrier within SM.
    - texture caches, thread dispatch logic etc.

Typical Cluster Set Up



JLab 10G cluster

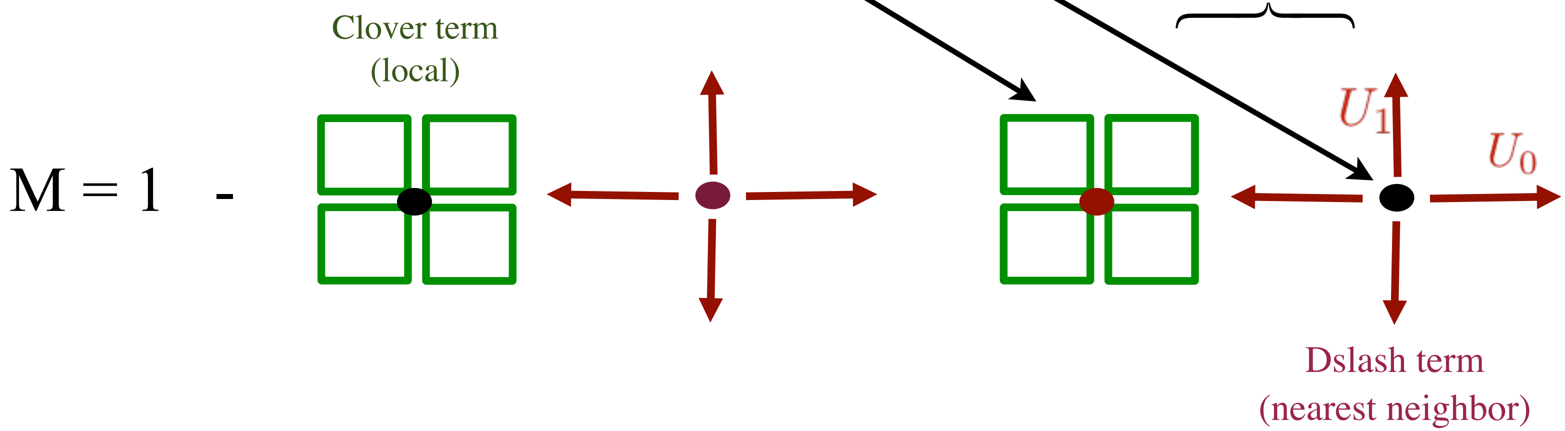
- GPU Mem. B/W / CPU Mem. B/W ~6.9x
- GPU Peak Flops (SP) / CPU Peak Flops(SP) ~ 8.4x
- PCIe Gen2 serious bottleneck for multi-GPU
- Balance will change with generations (core-i7, PCIe3,Kepler, FDR)
- JLab configuration: 4 GPUs, 2x4 core CPUs

The Wilson-Clover Fermion Matrix

After even-odd (red-black) preconditioning (Schur style):

$$M = 1 - A_{oo}^{-1} D_{oe} A_{ee}^{-1} D_{eo}$$

total: 1824 flops,
408 words in + 24 words out
FLOP/Byte: 1.06 (SP), 0.53 (DP)



- Rough 'Speed of Light' estimates - assuming streaming from memory
  - SU(3) Mv multiply/add imbalance: ~83% of peak Flops (Dslash)
  - bandwidth constraint: ~ 1x Mem B/W in Flops (SP) 0.5x (DP)
  - staggered is harder: ~(2/3) x Mem B/W in Flops (SP) 1/3x (DP)

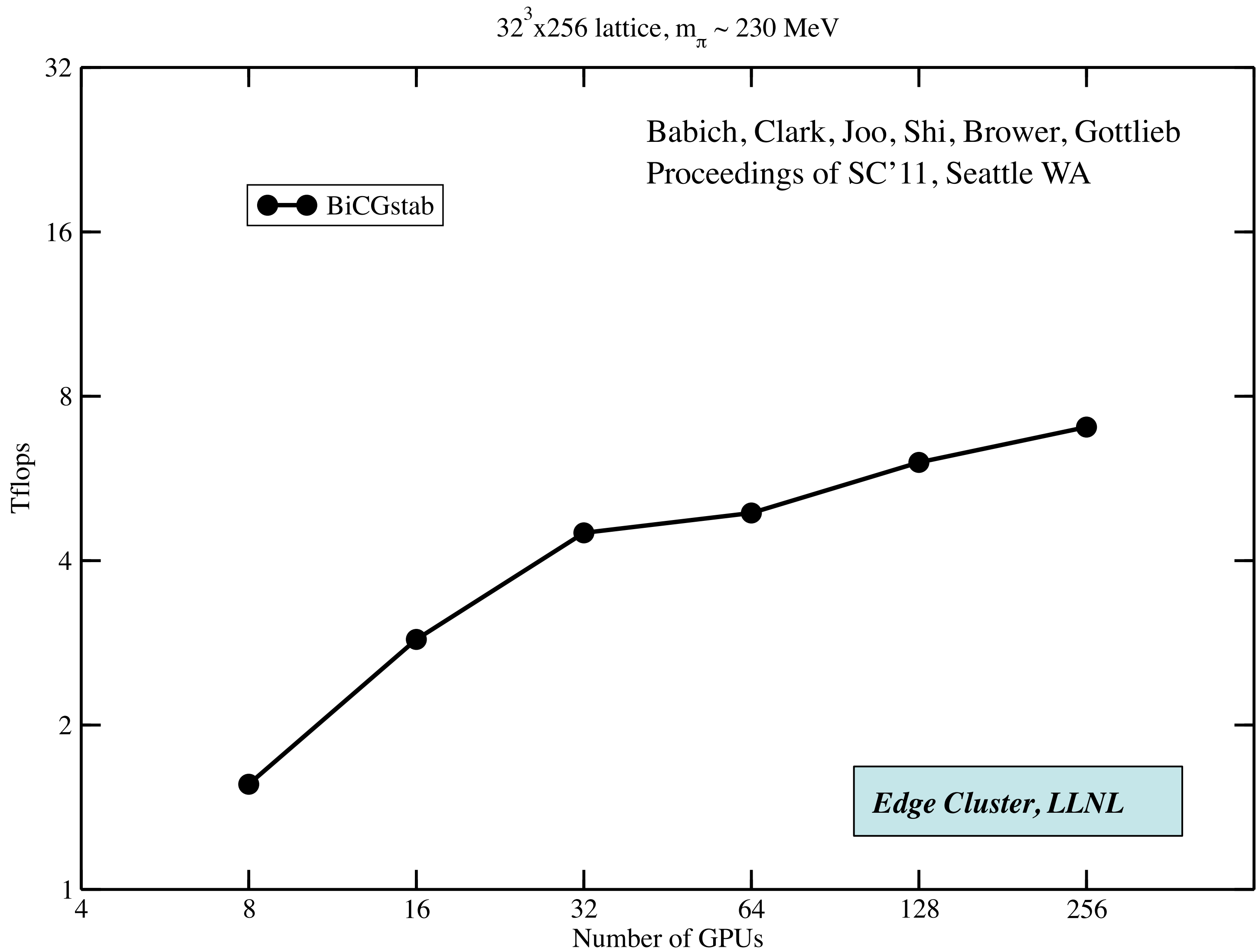
Enter QUDA

- QUDA is a library of solvers for lattice QCD on CUDA GPUs
  - *Clark, et. al., Comp. Phys. Commun. 181:1517-1528, 2010*
  - Supports: Wilson-Clover, Improved Staggered fermions
  - Domain Wall fermion support is ‘in development’
  - ‘Standard’ Krylov Solvers for QCD: CG(NE), BiCGStab
- Key Optimizations
  - Memory Bandwidth reducing techniques
    - Memory Coalescing Friendly Data Layout
    - Mixed Precision (16 bit, 32 bit, 64 bit) solvers
    - Field Compression
    - Dirac Basis (save loading half of t-neighbours)
    - Solve in Axial Gauge (save loading t-links)

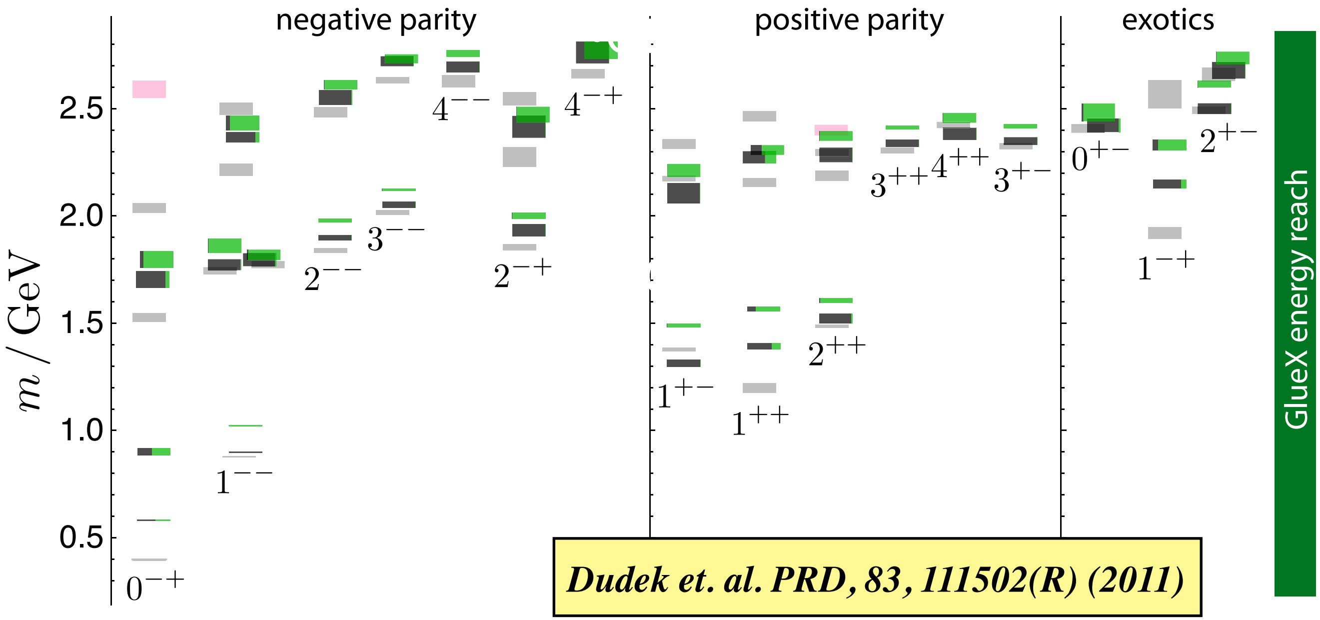
QUDA Community

- Integrated with Application Codes: Chroma & MILC
  - Enables production GPU use, by non GPU specialist scientists
  - Enlarges user base
- A group of interested developers coalesced around QUDA
  - Mike Clark (NVIDIA), Ron Babich (NVIDIA) - QUDA leads
  - Bálint Joó (Jefferson Lab) - Chroma integration
  - Guochun Shi (NCSA), Justin Foley (U. Utah) - MILC integration
  - Will Detmold, Joel Giedt, Alexei Strelchenko, Frank Winter, Chris Schroeder, Rich Brower, Steve Gottlieb
- Source Code Openly available from GitHub
  - http://github.com/lattice/quda

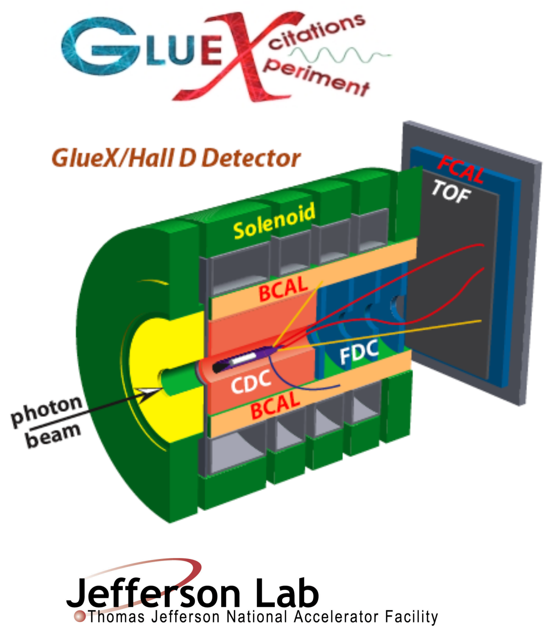
Solver Strong Scaling



Capacity Computing on GPUs



Dudek et. al. PRD, 83, 111502(R) (2011)

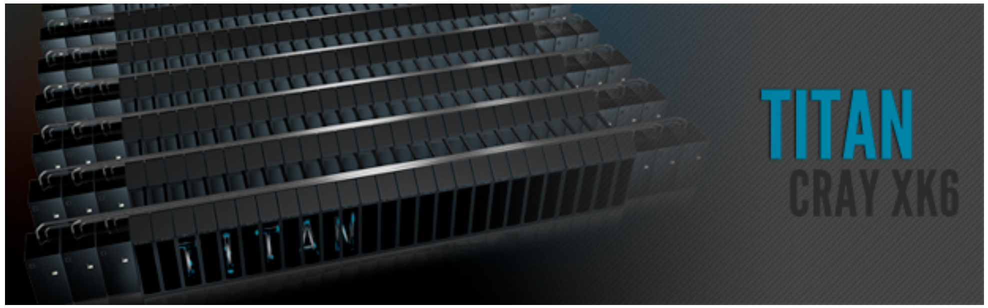


- **Capacity** (or high throughput) computing with small partitions (4-32 GPUs) is ideal for cost effective analysis
- Calculation of meson spectrum above:
  - 31 Million solves + variational basis + anisotropic lattices
  - Cost: about 1 month on USQCD National Facility GPU cluster at JLab. Currently around 500 GPUs in production use.
  - Exotics within reach of GlueX experiment of JLab @ 12GeV

Very Large Scale GPU machines

- Are already with us
  - Tianhe-1A
    - #1 on Top500 list Nov'10
  - Cray XK Architecture
    - OLCF Titan
    - NCSA BlueWaters
  - Large Clusters
    - Keeneland (NICS/NSF)
    - Edge (LLNL)
    - LOEWE-CSC (Frankfurt)
- Still hostage to PCIe
- Can QCD use such large systems 'at scale' ?

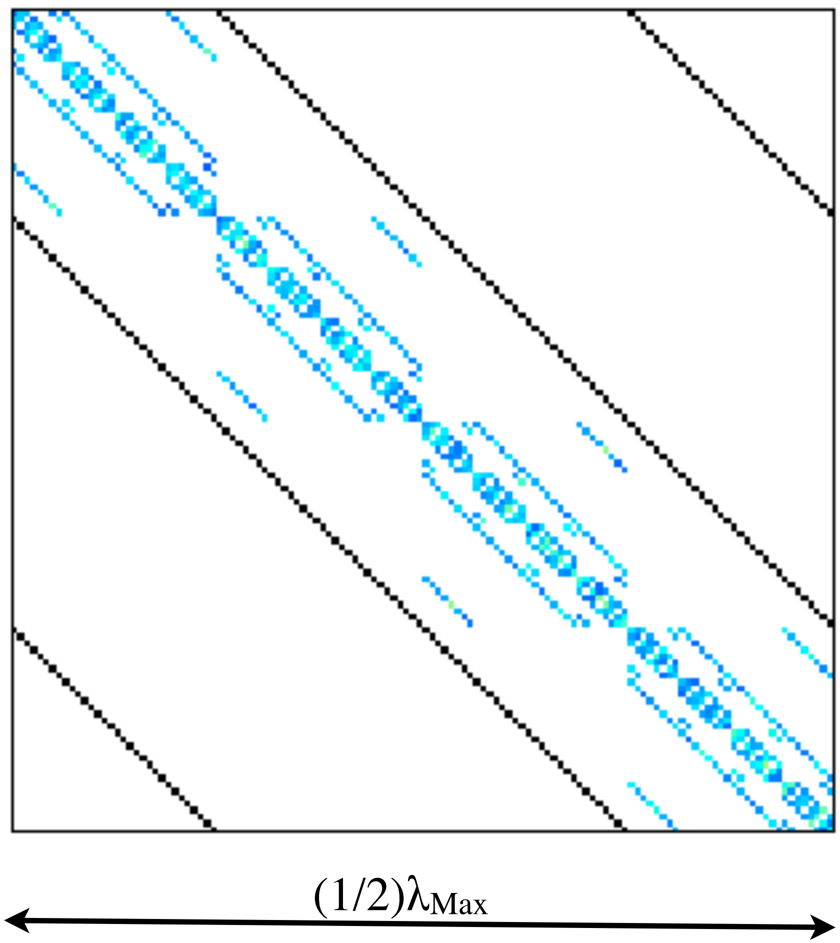
Tianhe-1A, National Supercomputing Center in Tianjin, China



Rendering of the forthcoming Titan Cray XK system at the Oak Ridge Leadership Computing Facility, Oak Ridge, TN, USA.

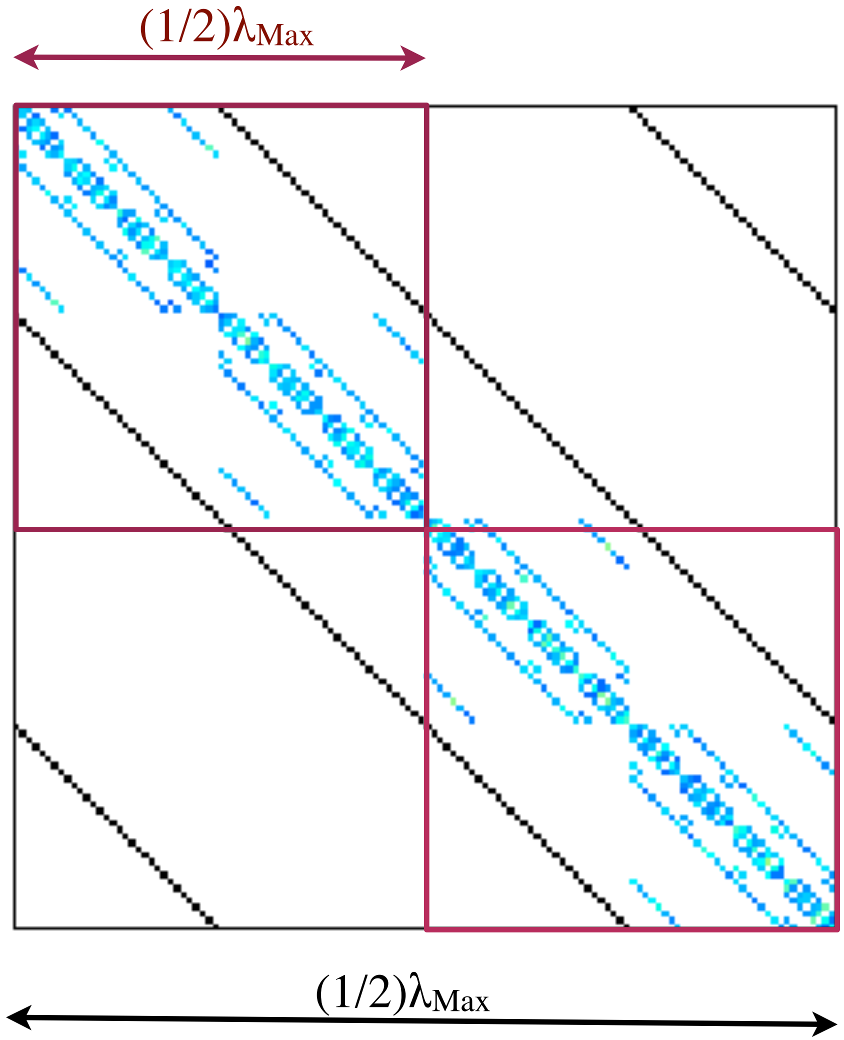
Keeneland NSF cluster Housed at NICS in Oak Ridge National Laboratory.

Reduced Communications Algorithm



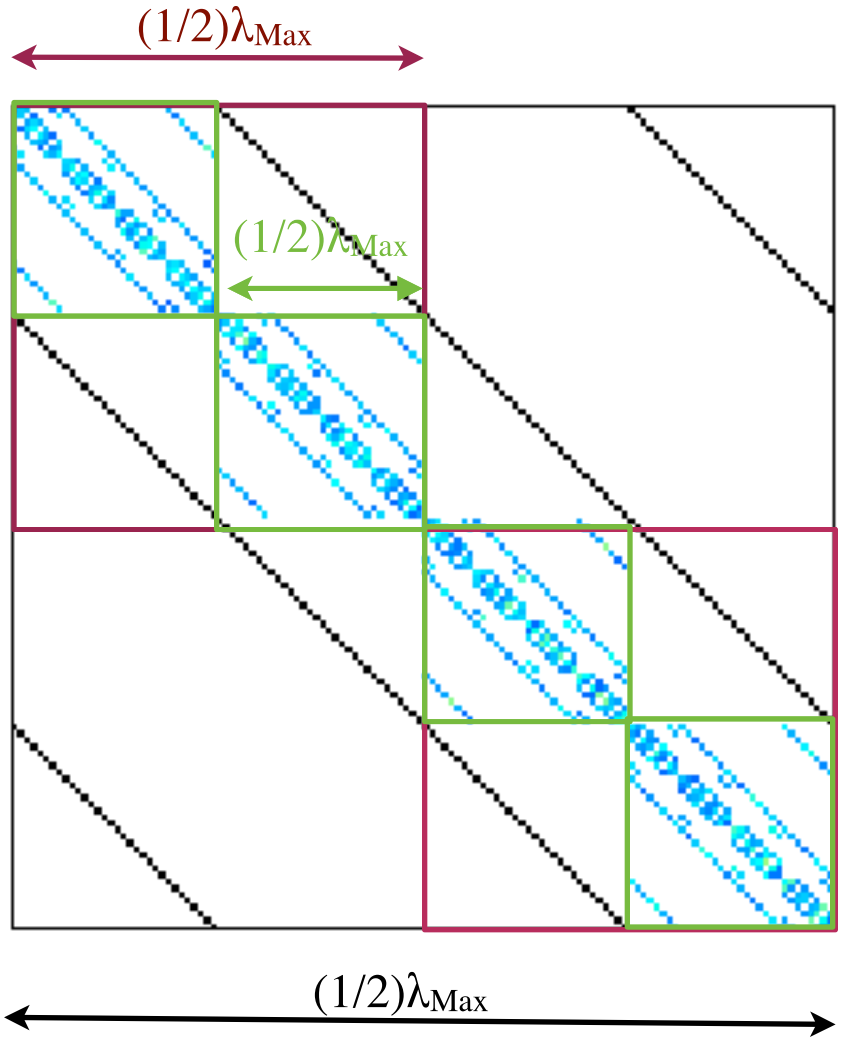
- Reduce Communication -> improve scaling
- Inner Block Diagonal Preconditioning solve
  - No communication between blocks
  - Can use reduced precision
- Outer Solver Process (GCR)
  - GCR needed to accommodate variable preconditioner.
- Blocks impose λ cutoff
  - Need to tune block size
- Heuristically (& from Lüscher)
  - keep wavelengths of ~ $O(\Lambda_{QCD}^{-1})$
  - $\Lambda_{QCD}^{-1} \sim 1$fm
  - Aniso: (a_s=0.125fm, a_t=0.035fm)
    - Our case: 8^3x32 blocks are ideal
  - Iso: 1fm ~ 8-10 sites (a=0.11fm)
  - Min. blocksize has scaling implications

Reduced Communications Algorithm



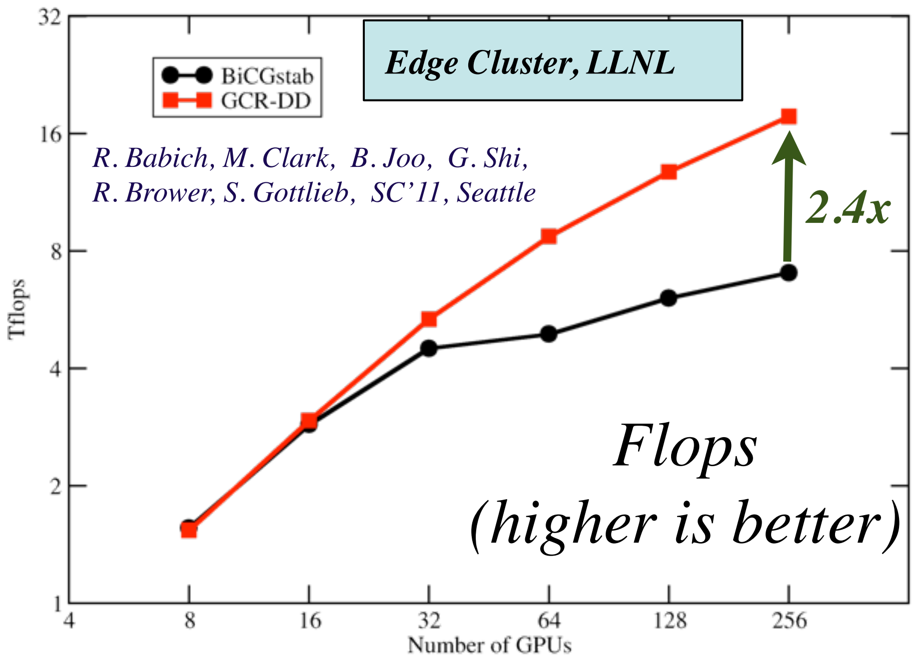
- Reduce Communication -> improve scaling
- Inner Block Diagonal Preconditioning solve
  - No communication between blocks
  - Can use reduced precision
- Outer Solver Process (GCR)
  - GCR needed to accommodate variable preconditioner.
- Blocks impose λ cutoff
  - Need to tune block size
- Heuristically (& from Lüscher)
  - keep wavelengths of ~ $O(\Lambda_{QCD}^{-1})$
  - $\Lambda_{QCD}^{-1} \sim 1$fm
  - Aniso: (a_s=0.125fm, a_t=0.035fm)
    - Our case: 8^3x32 blocks are ideal
  - Iso: 1fm ~ 8-10 sites (a=0.11fm)
  - Min. blocksize has scaling implications

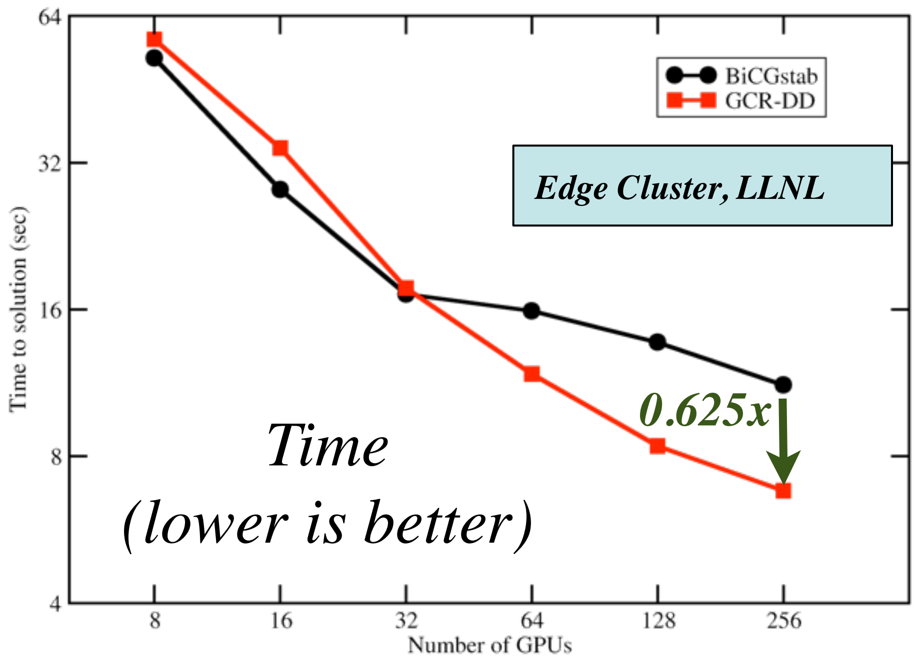
Reduced Communications Algorithm



- Reduce Communication -> improve scaling
- Inner Block Diagonal Preconditioning solve
  - No communication between blocks
  - Can use reduced precision
- Outer Solver Process (GCR)
  - GCR needed to accommodate variable preconditioner.
- Blocks impose λ cutoff
  - Need to tune block size
- Heuristically (& from Lüscher)
  - keep wavelengths of ~ $O(\Lambda_{QCD}^{-1})$
  - $\Lambda_{QCD}^{-1} \sim$ 1fm
  - Aniso: (a_s=0.125fm, a_t=0.035fm)
    - Our case: 8^3x32 blocks are ideal
  - Iso: 1fm ~ 8-10 sites (a=0.11fm)
  - Min. blocksize has scaling implications

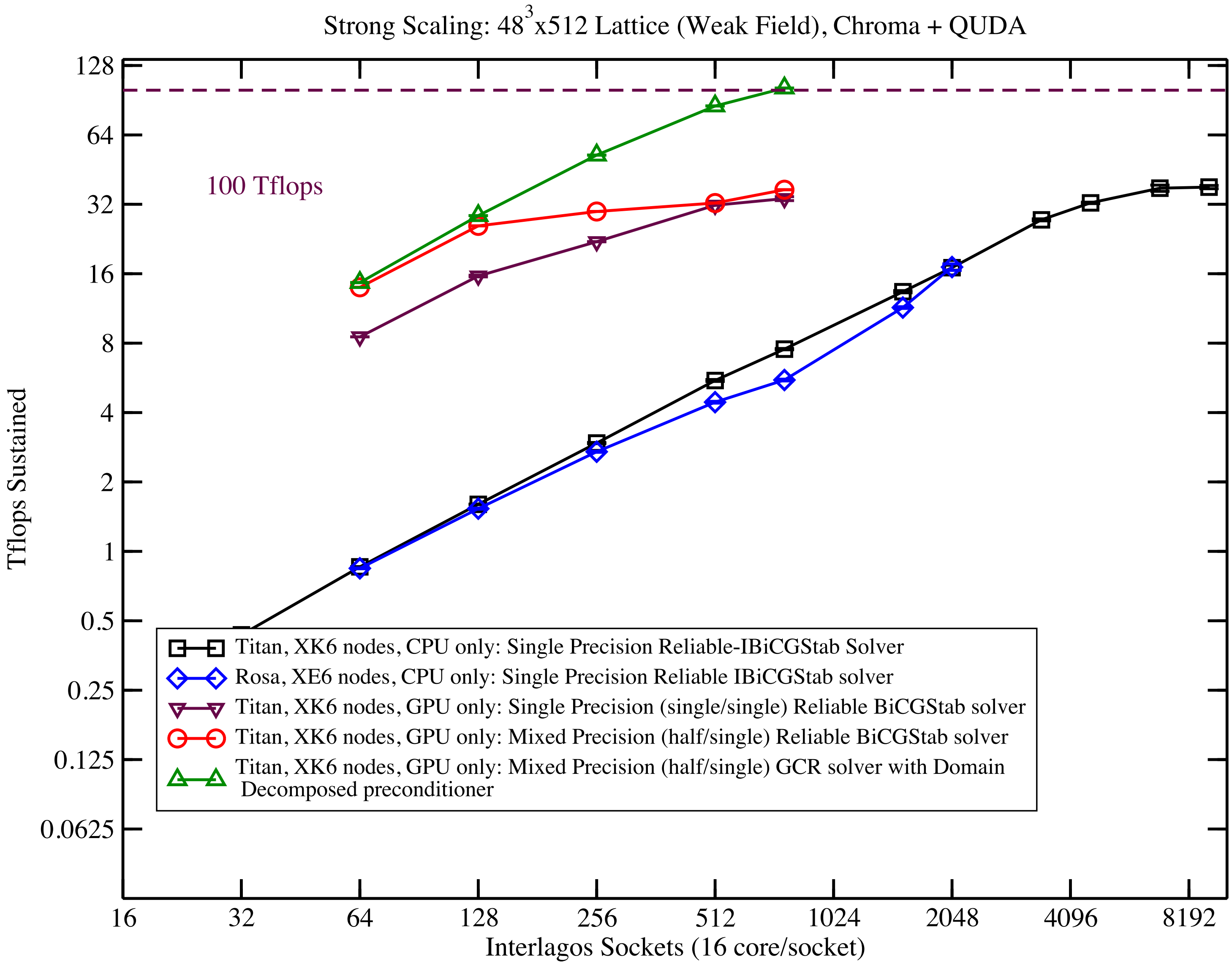
Scaling of DD+GCR vs BiCGStab



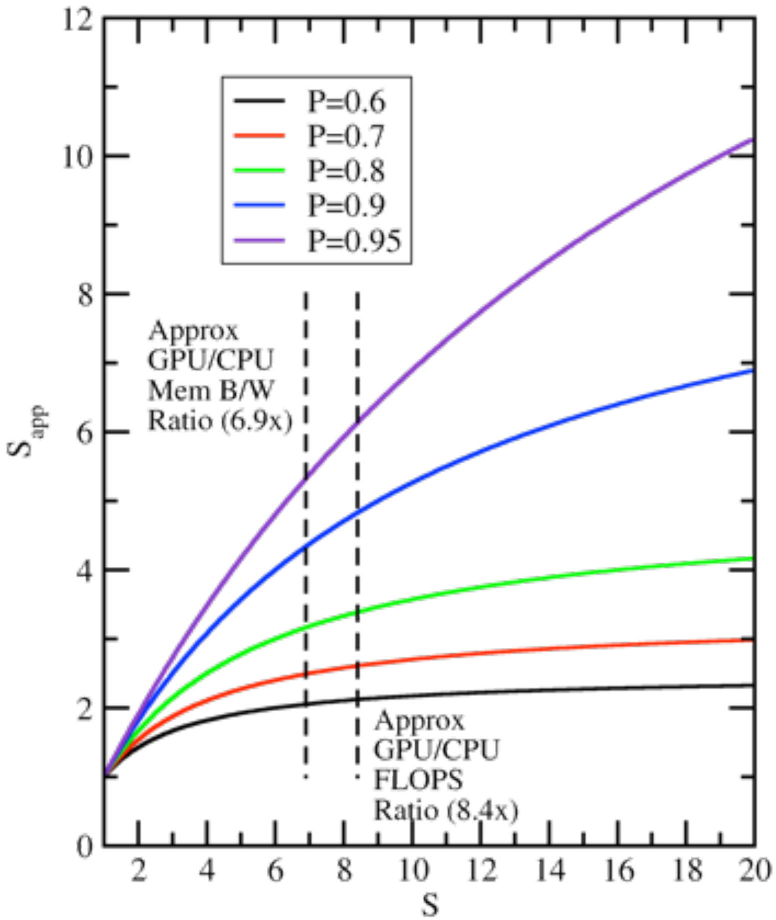


- SC'11 result (Edge Cluster, LLNL), 32^3x256 production lattices
- DD+GCR gets 2.4x BiCGStab flops, but only 1.6x gain in wall-time
  - Conservative factor: 1 DD-GCR flop ~ 1.5 BiCGStab flop
  - but factor is probably volume and partition size dependent also

Very recent results from TitanDev



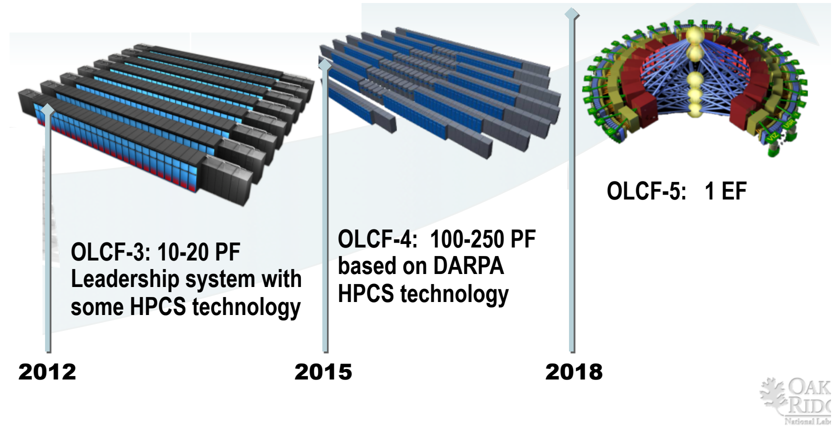
Beating Down Amdahl's Law

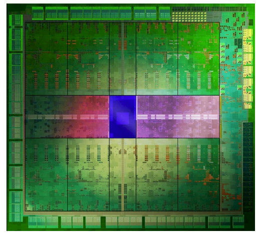


- "Distillation" technique spends 95% in solver
  - Perfect for GPUs, Very expensive otherwise
- Gauge Generation and Analysis Contractions are less solver bound
  - Gauge Generation: MD-forces (outside of solver)
  - Contractions: Lots of sums/inner products
- Need to move non-solver code to Accelerators
- Work in progress: Just-In-Time Compilation of QDP++ expressions on accelerators
  - In collaboration with F. Winter, University of Edinburgh
  - Gauge Generation Testing: B. Joo & F. Winter using Titan-Dev at OLCF
  - Analysis Testing: R. G. Edwards, & F. Winter, using JLab resources
  - See also: F. Winter "Accelerating QDP++ using GPUs" arXiv:1105:2279[hep-lat]

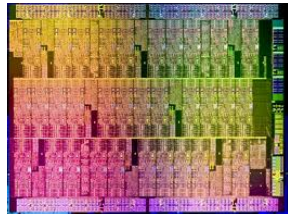
Future Architectures

- Foreseeable Leadership Computing Architectures
  - Cray XK series (Cray/NVIDIA) e.g. Titan
  - Stampede (Intel MIC)
  - BlueGene/Q (other swim lane)
  - Large scale GPU clusters
- Will GPUs remain GPUs?
  - Integration of GPU & CPU
  - Already in mobile/embedded
    - power efficiency = better battery life
    - Llano, Tegra, Intel Ivy Bridge

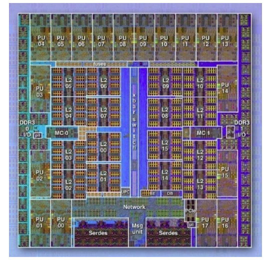




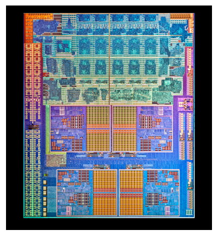
NVIDIA Kepler (1) (The Register)



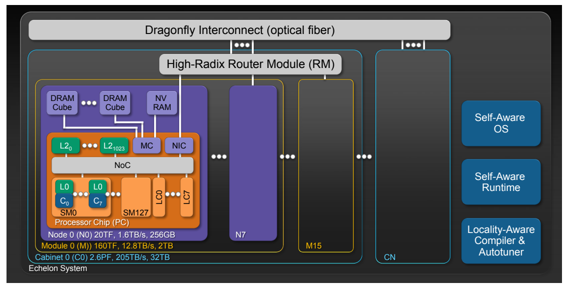
Intel MIC architecture (techeta.com)



IBM BG/Q Die (HPCWire)



AMD Llano



NVIDIA Echelon Design (SC'10)

Conclusions

- Lattice QCD was an early adopter of GPU technology
- Codes using the QUDA library can successfully use GPUs for science
  - GPU Clusters (capacity mode)
  - Large Scale GPU based resources (capability mode)
- GPUs enabled use of the “distillation” technique for analysis
- Scaling of ‘brute force multi-GPU’ codes is limited
  - communications (GPU->host->MPI->host->GPU) bottleneck
  - Architecture aware solvers can scale to 100s (possibly 1000s) of GPUS
- Large scale GPU machines ***are*** ~~coming soon~~ in a centre near you.
- Work is underway to port all of QDP++ to GPUs (QDP-JIT)
- Also exploring other architectures such as Intel MIC, BG/Q.
  - JLab is part of Intel MIC Software Development Program