# **NESA**

#### **GPU Clusters for HPC**

Volodymyr Kindratenko Innovative Systems Laboratory

National Center for Supercomputing Applications University of Illinois at Urbana-Champaign



### Acknowledgements

- ISL research staff
  - Jeremy Enos
  - Guochun Shi
  - Michael Showerman
  - Craig Steffen
  - Alexey Titov
- UIUC collaborators
  - Wen-Mei Hwu (ECE)
- Funding sources
  - NSF grants 0810563 and 0626354
  - NASA grant NNG06GH15G
  - NARA/NSF grant
  - IACAT's Center for Extreme-scale Computing





#### **ISL Research**

- Evaluation of emerging computing architectures
  - Reconfigurable computing
  - Many-core (GPU) architecture
  - Heterogeneous clusters
- Systems software research and development
  - Run-time systems
  - GPU accelerator cluster management
    - Tools and utilities: GPU memory test, power profiling, etc.
- Application development for emerging computing architectures
  - Computational chemistry (electronic structure, MD)
  - Computational physics (QCD)
  - Cosmology
  - Data mining



### Top 10 from TOP-500

#### Green500

Rank	Site	Computer
1	National Supercomputing Center in Tianjin	Tianhe-1A - NUDT TH MPP, X5670 2.93Ghz 6C, <mark>NVIDIA</mark> GPU, FT-1000 8C
2	DOE/SC/Oak Ridge National Laboratory	Jaguar - Cray XT5-HE Opteron 6-core 2.6 GHz
3	National Supercomputing Centre in Shenzhen (NSCS)	Nebulae - Dawning TC3600 Blade, Intel X5650, NVIDIA Tesla C2050 GPU
4	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.0 - HP ProLiant SL390s G7 Xeon 6C X5670, NVIDIA GPU, Linux/Windows
5	DOE/SC/LBNL/NERSC	Hopper - Cray XE6 12-core 2.1 GHz
6	Commissariat a l'Energie Atomique (CEA)	Tera-100 - Bull bullx super-node S6010/S6030
7	DOE/NNSA/LANL	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband
8	National Institute for Computational Sciences/University of Tennessee	Kraken XT5 - Cray XT5-HE Opteron 6-core 2.6 GHz
9	Forschungszentrum Juelich (FZJ)	JUGENE - Blue Gene/P Solution
10	DOE/NNSA/LANL/SNL	Cielo - Cray XE6 8-core 2.4 GHz

Rank	Site	Computer
1	IBM Thomas J. Watson Research Center	NNSA/SC Blue Gene/Q Prototype
2	GSIC Center, Tokyo Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5670, NVIDIA GPU, Linux/Windows
3	NCSA	Hybrid Cluster Core i3 2.93Ghz Dual Core, NVIDIA C2050, Infiniband
4	RIKEN Advanced Institute for Computational Science	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect
5	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D- Torus
5	Universitaet Regensburg	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D- Torus
5	Universitaet Wuppertal	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D- Torus
8	Universitaet Frankfurt	Supermicro Cluster, QC Opteron 2.1 GHz, ATI Radeon GPU, Infiniband
9	Georgia Institute of Technology	HP ProLiant SL390s G7 Xeon 6C X5660 2.8Ghz, NVIDIA Fermi, Infiniband QDR
10	National Institute for Environmental Studies	GOSAT Research Computation Facility, NVIDIA



### **QP: first GPU cluster** at NCSA

- 16 HP xw9400 workstations
  - 2216 AMD Opteron 2.4 GHz dual socket dual core
  - 8 GB DDR2
  - PCI-E 1.0
  - Infiniband QDR
- 32 Quadro Plex Computing Servers
  - 2 Quadro FX 5600 GPUs
  - 2x1.5 GB GDDR3
  - 2 per host





# Lincoln: First GPU-based TeraGrid production system

- Dell PowerEdge 1955 server
  - Intel 64 (Harpertown) 2.33
     GHz dual socket quad-core
  - 16 GB DDR2
  - Infiniband SDR
- Tesla S1070 1U GPU Computing Server
  - 1.3 GHz Tesla T10 processors
  - 4x4 GB GDDR3 SDRAM
- Cluster
  - Servers: 192
  - Accelerator Units: 96





# **QP follow-up: AC**





#### AC01-32 nodes

- HP xw9400 workstation
  - 2216 AMD Opteron 2.4 GHz dual socket dual core
  - 8GB DDR2 in ac04-ac32
  - 16GB DDR2 in ac01-03,
     "bigmem" on qsub line
  - PCI-E 1.0
  - Infiniband QDR
- Tesla S1070 1U GPU Computing Server
  - 1.3 GHz Tesla T10 processors
  - 4x4 GB GDDR3
  - 1 per host



DRAM

DRAM

 $\bigcirc$ 

DRAM

DRAM

Tesla S1070



#### Lincoln vs. AC: HPL Benchmark





#### AC34-AC41 nodes

- Supermicro A+ Server
  - Dual AMD 6 core Istanbul
  - 32 GB DDR2
  - PCI-E 2.0
  - QDR IB (32 Gbit/sec)
  - 3 Internal ATI Radeon 5870 GPUs





#### AC33 node

- CPU cores (Intel core i7): 8
- Accelerator Units (S1070): 2
- Total GPUs: 8
- Host Memory: 24-32 GB DDR3
- GPU Memory: 32 GB
- CPU cores/GPU ratio: 1:2
- PCI-E 2.0
- Dual IOH (72 lanes PCI-E)





Core i7 host



#### **AC42**

#### • TYAN FT72-B7015

- X5680 Intel Xeon 3.33 GHz (Westmere-EP) dual-sosket hexa-core
- Tylersburg-36D IOH
- 24 GB DDR3
- 8 PCI-E 2.0 ports
  - switched
- NVIDIA GTX 480
  - 480 cores
  - 1.5 GB GDDR5







#### **EcoG: #3 on Green500 list**





#### EcoG nodes designed for low power

- EVGA P55V 120-LF-E651-TR Micro ATX Intel Motherboard
  - Core i3 530 2.93 GHz single-socket dual-core
  - 4 GB DDR3
  - PCIe x16 Gen2
  - QDR Infiniband
- Tesla C2050
  - 448 cores
  - 3 GB GDDR5
- Cluster
  - 128 nodes







#### **GPU Cluster Software**

- Shared system software
  - Torque / Moab
  - ssh
- Programming tools
  - CUDA C SDK
  - OpenCL SDK
  - PGI+GPU compiler
  - Matlab
  - Intel compiler
- Other tools
  - mvapich2 mpi (IB)

- Unique to AC
  - CUDA wrapper
  - memtest
  - Power profiling



#### **Need for GPU-aware cluster software stack**

- Issues unique to compute nodes with GPUs
  - Thread affinity mapping to maximize host-GPU bandwidth
  - CUDA API/driver software bugs (mainly in initial product releases)
  - ...
- Issues unique to the GPU cluster
  - Efficient GPUs sharing in a multi-user environment
  - GPU memory cleanup between different users
  - ...
- Other issues of interest
  - Are GPUs reliable? (Non-ECC memory in initial products)
  - Are they power-efficient?
  - •



# **Effects of NUMA**

• On some systems

access time from CPU00 to GPU0 ≠ access time from CPU10 to GPU0

- Latency and achievable bandwidth are affected
- Solution: automatic affinity mapping for user processes depending on the GPUs used





#### **Host to device Bandwidth Comparison**





#### **Efficient GPU resources sharing**

- In a typical cluster environment, user obtains exclusive access to the entire node
- A typical GPU cluster node has few (2-4) GPUs
- But a typical GPU cluster application is designed to utilize only one GPU per node
  - Giving access to the entire cluster node is wasteful if the user only needs one GPU
- Solution: allow user to specify how many GPUs his application needs and fence the remaining GPUs for other users





## **CUDA/OpenCL Wrapper**

- Basic operation principle:
  - Use /etc/ld.so.preload to overload (intercept) a subset of CUDA/OpenCL functions, e.g. {cu|cuda}{Get|Set}Device, clGetDeviceIDs, etc.
  - Transparent operation
- Purpose:
  - Enables controlled GPU device visibility and access, extending resource allocation to the workload manager
  - Prove or disprove feature usefulness, with the hope of eventual uptake or reimplementation of proven features by the vendor
  - Provides a platform for rapid implementation and testing of HPC relevant features not available in NVIDIA APIs
- Features:
  - NUMA Affinity mapping
    - Sets thread affinity to CPU core(s) nearest the gpu device
  - Shared host, multi-gpu device fencing
    - Only GPUs allocated by scheduler are visible or accessible to user
    - GPU device numbers are virtualized, with a fixed mapping to a physical device per user environment
    - User always sees allocated GPU devices indexed from 0



# **CUDA/OpenCL Wrapper**

- Features (cont'd):
  - Device Rotation (deprecated)
    - Virtual to Physical device mapping rotated for each process accessing a GPU device
    - Allowed for common execution parameters (e.g. Target gpu0 with 4 processes, each one gets separate gpu, assuming 4 gpus available)
    - CUDA 2.2 introduced *compute-exclusive* device mode, which includes fallback to next device. Device rotation feature may no longer needed.
  - Memory Scrubber
    - Independent utility from wrapper, but packaged with it
    - Linux kernel does no management of GPU device memory
    - Must run between user jobs to ensure security between users
- Availability
  - NCSA/Uofl Open Source License
  - https://sourceforge.net/projects/cudawrapper/



#### wrapper\_query utility

• Within any job environment, get details on what the wrapper library is doing

💭 2:ac - default* - SSH Secure Shell 📃 🖉	5
Eile Edit View Window Help	
🛛 🖶 🎒 🕼 🖀 🖹 🙀 🎒 🆓 🆓 🦑 🎌	
[jenos@ac21 ~]\$ wrapper query	*
cuda_wrapper info:	
version=2	
userID=20152	
pid=-1	
nGPU=4	
physGPU[0]=0	
physGPU[1]=1	
physGPU[2]=2	
physGPU[3]=3	
key_env_var=	
allow_user_passthru=1	
affinity:	
GPU=0, CPU=0 2	
GPU=1, CPU=0 2	
GPU=2,  CPU=1 3	
GPU=3, CPU=1 3	
CudaAPI = Runtime API	
walltime = $12.902300$ seconds	
$gpu_kernet_crime = 44.494469$ seconds	
jonoglac21 mls	
Connected to ac SSH2 - aes128-cbc - hmac-n 80x22 🗱 NUM	1



#### showgputime

- Shows percent time CUDA linked processes utilized GPU
- Displays last 15 records (showallgputime shows all)
- Requires support of cuda\_wrapper implementation

BA ON SA							
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	🖗 N? 🛛 🛃 Quick	Connect 🛄 Profiles					
showqputime							
		+		+			+
job_id	job_name	node_name	cudaAPI	nGPU	wall_time	gpu_kernel_time	percent_gputime
532213.acm	gpu use test	ac22	Runtime API	1	12.247	11.1232	90.8
529419.acm	sanity check	ac33	Runtime API	4	86.6741	0	0.0
529419.acm	sanity check	ac03	Runtime API	4	40.649	44.4939	27.4
529419.acm	sanity check	ac02	Runtime API	4	52.863	0	0.0
529419.acm	sanity check	ac01	Runtime API	4	29.2561	0	0.0
529418.acm	sanity check	ac27	Unknown	4	0	0	NULL
529418.acm	sanity_check	ac23	Unknown	4	0	0	NULL
529418.acm	sanity check	ac19	Unknown	4	0	0	NULL
529418.acm	sanity check	ac17	Unknown	4	0	0	NULL
529417.acm	sanity check	ac33	Unknown	4	0	0	NULL
529417.acm	sanity check	ac03	Unknown	4	0	0	NULL
529417.acm	sanity_check	ac02	Unknown	4	0	0	NULL
529410.acm	sanity check	ac32	Runtime API	4	76.2256	88,9835	29.2
529410.acm	sanity_check	ac31	Runtime API	4	73.5471	88.9877	30.2
529410.acm	sanity_check	ac30	Runtime API	4	71.3086	88.9877	31.2
	job_id 532213.acm 529419.acm 529419.acm 529419.acm 529419.acm 529418.acm 529418.acm 529418.acm 529418.acm 529418.acm 529417.acm 529417.acm 529417.acm 529410.acm 529410.acm	job_id   job_name 532213.acm   gpu_use_test 529419.acm   sanity_check 529419.acm   sanity_check 529419.acm   sanity_check 529419.acm   sanity_check 529418.acm   sanity_check 529418.acm   sanity_check 529418.acm   sanity_check 529418.acm   sanity_check 529418.acm   sanity_check 529417.acm   sanity_check 529417.acm   sanity_check 529417.acm   sanity_check 529410.acm   sanity_check 529410.acm   sanity_check 529410.acm   sanity_check 529410.acm   sanity_check	job_id   job_name   node_name   532213.acm   gpu_use_test   ac22 529419.acm   sanity_check   ac33 529419.acm   sanity_check   ac03 529419.acm   sanity_check   ac02 529419.acm   sanity_check   ac01 529418.acm   sanity_check   ac27 529418.acm   sanity_check   ac23 529418.acm   sanity_check   ac19 529418.acm   sanity_check   ac19 529418.acm   sanity_check   ac17 529418.acm   sanity_check   ac13 529417.acm   sanity_check   ac03 529417.acm   sanity_check   ac03 529417.acm   sanity_check   ac03 529417.acm   sanity_check   ac03 529417.acm   sanity_check   ac32 529410.acm   sanity_check   ac31 529410.acm   sanity_check   ac30	job_id   job_name   node_name   cudaAPI 532213.acm   gpu_use_test   ac22   Runtime_API 529419.acm   sanity_check   ac33   Runtime_API 529419.acm   sanity_check   ac03   Runtime_API 529419.acm   sanity_check   ac02   Runtime_API 529419.acm   sanity_check   ac01   Runtime_API 529419.acm   sanity_check   ac27   Unknown 529418.acm   sanity_check   ac23   Unknown 529418.acm   sanity_check   ac19   Unknown 529418.acm   sanity_check   ac17   Unknown 529418.acm   sanity_check   ac13   Unknown 529418.acm   sanity_check   ac33   Unknown 529417.acm   sanity_check   ac03   Unknown 529417.acm   sanity_check   ac03   Unknown 529417.acm   sanity_check   ac03   Unknown 529417.acm   sanity_check   ac32   Runtime_API 529410.acm   sanity_check   ac31   Runtime_API 529410.acm   sanity_check   ac30   Runtime_API	job_id   job_name   node_name   cudaAPI   nGPU 532213.acm   gpu_use_test   ac22   Runtime_API   1 529419.acm   sanity_check   ac33   Runtime_API   4 529419.acm   sanity_check   ac03   Runtime_API   4 529419.acm   sanity_check   ac02   Runtime_API   4 529419.acm   sanity_check   ac01   Runtime_API   4 529419.acm   sanity_check   ac01   Runtime_API   4 529418.acm   sanity_check   ac27   Unknown   4 529418.acm   sanity_check   ac23   Unknown   4 529418.acm   sanity_check   ac19   Unknown   4 529418.acm   sanity_check   ac17   Unknown   4 529418.acm   sanity_check   ac17   Unknown   4 529417.acm   sanity_check   ac33   Unknown   4 529417.acm   sanity_check   ac03   Unknown   4 529417.acm   sanity_check   ac03   Unknown   4 529417.acm   sanity_check   ac03   Unknown   4 529410.acm   sanity_check   ac31   Runtime_API   4 529410.acm   sanity_check   ac30   Runtime_API   4 529410	job_id       job_name       node_name       cudaAPI       nGPU       wall_time         532213.acm       gpu_use_test       ac22       Runtime_API       1       12.247         529419.acm       sanity_check       ac33       Runtime_API       4       86.6741         529419.acm       sanity_check       ac03       Runtime_API       4       40.649         529419.acm       sanity_check       ac02       Runtime_API       4       29.2561         529419.acm       sanity_check       ac01       Runtime_API       4       29.2561         529418.acm       sanity_check       ac27       Unknown       4       0         529418.acm       sanity_check       ac23       Unknown       4       0         529418.acm       sanity_check       ac19       Unknown       4       0         529418.acm       sanity_check       ac17       Unknown       4       0         529417.acm       sanity_check       ac33       Unknown       4       0         529417.acm       sanity_check       ac03       Unknown       4       0         529417.acm       sanity_check       ac03       Unknown       4       0         529417.ac	job_id       job_name       node_name       cudaAPI       nGPU       wall_time       gpu_kernel_time         532213.acm       gpu_use_test       ac22       Runtime_API       1       12.247       11.1232         529419.acm       sanity_check       ac33       Runtime_API       4       86.6741       0         529419.acm       sanity_check       ac03       Runtime_API       4       46.6741       0         529419.acm       sanity_check       ac03       Runtime_API       4       46.6741       0         529419.acm       sanity_check       ac02       Runtime_API       4       52.863       0         529419.acm       sanity_check       ac01       Runtime_API       4       29.2561       0         529418.acm       sanity_check       ac23       Unknown       4       0       0         529418.acm       sanity_check       ac23       Unknown       4       0       0         529418.acm       sanity_check       ac17       Unknown       4       0       0         529417.acm       sanity_check       ac03       Unknown       4       0       0         529417.acm       sanity_check       ac03       Unknown

#### **Are GPUs reliable?**

- No ECC in initial product releases
  - Not a big deal when a GPU is used for what it was indented: image rendering
  - Could be a problem when executing a scientific application
    - Can we trust the computed results?
    - How do we know the results are correct?
- Fermi architecture now has ECC memory protection
- However, two years ago it was not clear if NVIDIA was going to add ECC
  - We have done a GPU memory reliability study



#### **CUDA Memtest**

- Features
  - Full re-implementation of every test included in memtest86
  - Random and fixed test patterns, error reports, error addresses, test specification
  - Includes additional stress test for software and hardware errors
  - Email notification
- Usage scenarios
  - Hardware test for defective GPU memory chips
  - CUDA API/driver software bugs detection
  - Hardware test for detecting soft errors due to non-ECC memory
  - Stress test for thermal loading
- No soft error detected in 2 years x 4 gig of cumulative runtime
- But several Tesla units in AC and Lincoln clusters were found to have hard memory errors (and thus have been replaced)
- Availability
  - NCSA/Uofl Open Source License
  - https://sourceforge.net/projects/cudagpumemtest/



#### **GPU Node Pre/Post Allocation Sequence**

- Pre-Job (minimized for rapid device acquisition)
  - Assemble detected device file unless it exists
  - Sanity check results
  - Checkout requested GPU devices from that file
  - Initialize CUDA wrapper shared memory segment with unique key for user (allows user to ssh to node outside of job environment and have same gpu devices visible)
- Post-Job
  - Use quick memtest run to verify healthy GPU state
  - If bad state detected, mark node offline if other jobs present on node
  - If no other jobs, reload kernel module to "heal" node (for CUDA driver bug)
  - Run memscrubber utility to clear gpu device memory
  - Notify of any failure events with job details via email
  - Terminate wrapper shared memory segment
  - Check-in GPUs back to global file of detected devices



#### **Are GPUs power-efficient?**

- GPUs are power-hungry
  - GTX 480 250 W
  - C2050 238 W
- But does the increased power consumption justify their use?
  - How much power do jobs use?
  - How much do they use for pure CPU jobs vs. GPU-accelerated jobs?
  - Do GPUs deliver a hoped-for improvement in power efficiency?
  - How do we measure actual power consumption?
  - How do we characterize power efficiency?



### **Power Profiling Tools**

#### Goals:

- Accurately record power consumption of GPU and workstation (CPU) for performance per watt efficiency comparison
- Make this data clearly and conveniently presented to application users
- Accomplish this with cost effective hardware

Solution:

- Modify inexpensive power meter to add logging capability
- Integrate monitoring with job management infrastructure
- Use web interface to present data in multiple forms to user



#### **Power Profiling Hardware**

Tweet-a-Watt



- Wireless receiver (USB)
- AC01 host and associated GPU unit are monitored separately by two Tweet-a-Watt transmitters
- Measurements are reported every 30 seconds
- < \$100 total parts</p>



#### **Power Profiling Walk Through**

- Submit job with prescribed resource (powermon)
- Run application as usual, follow link(s)

🗾 2:ac - default* - SSH Secure Shell	
Ele Edit View Window Help	
🖥 🗃 🕼 🗂 🖉 陷 🕾 🖻 🛤 🍰 🚰 🧐 🤣 🛠 😯 🔢 Quick Connect. 🚞 Profiles	
[jenos@ac ~]\$ qsub -I -l nodes=1:ppn=4:powermon qsub: waiting for job 532214.acm to start qsub: job 532214.acm ready	
This job is running on a power profile node. (experime View job power profile at: http://ac.ncsa.uiuc.edu/power.php?jobid=532214.acm Or compare to other jobs at: http://ac.ncsa.uiuc.edu/jobs.php	ntal feature)
[jenos@ac01 ~]\$	
Connected to ac SSH2 - aes	28-cbc - hmac-n 80x21 🛛 🖉 NUM



#### **Power Profiling Walk Through**

Elle Edit View Higtory Bookmarks Tools Help         AC Power Monitor       Inttp://ac.ncsa.uiuc.edu/jobs.php       Image: Comparison of the start of th	실 AC Power I	Monitor - Mozilla Firefox					_ 0	X
Image: Construction of the second	<u>F</u> ile <u>E</u> dit <u>V</u> i	iew Hi <u>s</u> tory <u>B</u> ookmarks <u>T</u> ools <u>H</u>	elp					
AC Power Monitor       Image: Start Time       Duration       Data         532214.acm       STDIN       jenos       2010-04.14 03:00:34       06.44       552214.acm.csv         532191.acm       coarse_mes       atorok       2010-04.13 23:43:31       125:20       552191.acm.csv         532171.acm       coarse_mes       atorok       2010-04.13 23:43:31       125:20       552191.acm.csv         532171.acm       coarse_mes       atorok       2010-04.13 20:45:36       127:58       552157.acm.csv         552137.acm       coarse_mes       atorok       2010-04.13 19:15:36       128:54       552137.acm.csv         552121.acm       coarse_mes       atorok       2010-04.13 19:15:36       128:54       552137.acm.csv         552105.acm       coarse_mes       atorok       2010-04.13 19:15:36       128:54       552137.acm.csv         552105.acm       coarse_mes       atorok       2010-04.13 18:21:00       127:12       552105.acm.csv         551199.acm       oarse_mes       atorok       2010-04.01 16:32:07       10:21       55054.acm.csv         550958.acm       STDIN       gshi       2010-04.07 16:38:07       40:31:35       550958.acm.csv         550510.acm       coarse_mes       atorok       2010-04.02 20:95:5	🤙 🧼 ⊙ 🤅	🥭 🕘 😭 🗈 http://ac.ncs	a.uiuc.edu	/jobs.php 🛞	10 : 😗	• Google		P
AC power-monitored jobs         Chart       Job Name       user       Start Time       Duration       Data         532214.acm       STDIN       jenos       2010-04-14 03:00:34       06:44       532214.acm.csv         532191.acm       coarse_mes       atorok       2010-04-13 23:43:31       1:25:00       532191.acm.csv         532171.acm       coarse_mes       atorok       2010-04-13 22:14:47       1:27:30       53217.acm.csv         532173.acm       coarse_mes       atorok       2010-04-13 20:45:36       1:27:58       53215.acm.csv         532173.acm       coarse_mes       atorok       2010-04-13 12:14:01       1:27:12       53217.acm.csv         532173.acm       coarse_mes       atorok       2010-04-13 12:04:536       1:27:12       53215.acm.csv         532105.acm       coarse_mes       atorok       2010-04-13 17:49:22       1:24:56       532121.acm.csv         532105.acm       coarse_mes       atorok       2010-04-13 16:21:00       1:27:12       532105.acm.csv         531876.acm       coarse_mes       atorok       2010-04-07 16:58:07       40:31:35       500958.acm.csv         530958.acm       STDIN       gshi       2010-04-07 16:32:47       1:02:1       530610.acm.csv         530050.acm<	AC Power	Monitor						$\odot$
Chart         Job Name         user         Start Time         Duration         Data           532214.acm         STDIN         jenos         2010-04-14 03:00:34         06:44         532214.acm csv           532191.acm         coarse_mes         atorok         2010-04-13 23:43:31         1:25:20         532191.acm csv           532171.acm         coarse_mes         atorok         2010-04-13 23:43:31         1:25:20         532171.acm csv           532173.acm         coarse_mes         atorok         2010-04-13 20:45:36         1:27:58         532155.acm csv           532137.acm         coarse_mes         atorok         2010-04-13 19:15:36         1:28:54         532137.acm csv           532105.acm         coarse_mes         atorok         2010-04-13 16:21:00         1:27:12         53216.acm csv           532105.acm         coarse_mes         atorok         2010-04-13 16:21:00         1:27:12         53216.acm csv           531876.acm         coarse_mes         atorok         2010-04-16:38:07         40:31:35         530958.acm csv           530958.acm         STDIN         gshi         2010-04-07 16:32:47         0:10:12         530958.acm csv           530610.acm         coarse_mes         atorok         2010-04-06 23:40:12         1:27:11         <	AC power-	monitored jobs						ŕ
532214.acm       STDIN       jenos       2010-04-14 03:00:34       0:6:44       532214.acm.csv         532191.acm       coarse_mes       atorok       2010-04-13 23:43:31       1:25:20       532191.acm.csv         532171.acm       coarse_mes       atorok       2010-04-13 22:14:47       1:27:30       532171.acm.csv         532155.acm       coarse_mes       atorok       2010-04-13 20:45:36       1:27:58       532137.acm.csv         532137.acm       coarse_mes       atorok       2010-04-13 19:15:36       1:28:54       532137.acm.csv         532121.acm       coarse_mes       atorok       2010-04-13 17:49:22       1:24:56       532121.acm.csv         532155.acm       coarse_mes       atorok       2010-04-13 16:21:00       1:27:12       532105.acm.csv         532165.acm       coarse_mes       atorok       2010-04-13 16:21:00       1:27:12       532105.acm.csv         531876.acm       coarse_mes       atorok       2010-04-13 16:21:00       1:27:12       53015.acm.csv         530958.acm       STDIN       gshi       2010-04-07 16:32:47       0:10:21       530954.acm.csv         530610.acm       coarse_mes       atorok       2010-04-06 23:40:12       1:27:1       530610.acm.csv         530562.acm       coarse_mes <th>Chart</th> <th>Job Name</th> <th>user</th> <th>Start Time</th> <th>Duration</th> <th>Data</th> <th></th> <th></th>	Chart	Job Name	user	Start Time	Duration	Data		
\$32191.acm       coarse_mes       atorok       2010-04-13 23:43:31       1:25:20       \$32191.acm.csy         \$32171.acm       coarse_mes       atorok       2010-04-13 20:41:47       1:27:30       \$32155.acm.csy         \$32155.acm       coarse_mes       atorok       2010-04-13 20:45:36       1:27:58       \$32155.acm.csy         \$32137.acm       coarse_mes       atorok       2010-04-13 19:15:36       1:28:54       \$32137.acm.csy         \$32121.acm       coarse_mes       atorok       2010-04-13 19:15:36       1:28:54       \$32121.acm.csy         \$32105.acm       coarse_mes       atorok       2010-04-13 16:21:00       1:27:12       \$32105.acm.csy         \$32105.acm       coarse_mes       atorok       2010-04-12 18:38:50       2:00       \$31876.acm.csy         \$31199.acm       STDIN       gshi       2010-04-07 16:58:07       40:31:35       \$30958.acm.csy         \$30054.acm       STDIN       gshi       2010-04-07 16:32:47       1:02:1       \$30010.acm.csy         \$300552.acm       coarse_mes       atorok       2010-04-06 23:40:12       1:27:1       \$30610.acm.csy         \$30050.acm       coarse_mes       atorok       2010-04-06 20:39:54       1:28:37       \$30562.acm.csy         \$30050.acm       coarse_mes	532214.acm	STDIN	jenos	2010-04-14 03:00:34	0:6:44	532214.acm.csv		
532171.acm       coarse_mes       atorok       2010-04-13 22:14:47       1:27:30       532171.acm.csv         532155.acm       coarse_mes       atorok       2010-04-13 20:45:36       1:27:58       532155.acm.csv         532137.acm       coarse_mes       atorok       2010-04-13 19:15:36       1:28:54       532137.acm.csv         532121.acm       coarse_mes       atorok       2010-04-13 19:15:36       1:28:54       532121.acm.csv         532105.acm       coarse_mes       atorok       2010-04-13 19:15:36       1:28:54       532121.acm.csv         532105.acm       coarse_mes       atorok       2010-04-13 16:21:00       1:27:12       532105.acm.csv         531876.acm       coarse_mes       atorok       2010-04-13 18:8:50       2:0.0       531876.acm.csv         531199.acm       STDIN       gshi       2010-04-07 16:32:47       0:10:21       530954.acm.csv         530958.acm       STDIN       gshi       2010-04-07 16:32:47       0:10:21       530954.acm.csv         530610.acm       coarse_mes       atorok       2010-04-07 16:32:47       0:10:21       530954.acm.csv         530585.acm       coarse_mes       atorok       2010-04-06 20:39:55       1:28:56       530585.acm.csv         530510.acm       coarse_mes	532191.acm	coarse_mes	atorok	2010-04-13 23:43:31	1:25:20	532191.acm.csv		
532155 acm       coarse_mes       atorok       2010-04-13 20:45:36       1:27:58       532155 acm.csv         532137.acm       coarse_mes       atorok       2010-04-13 19:15:36       1:28:54       532121.acm.csv         532105 acm       coarse_mes       atorok       2010-04-13 17:49:22       1:24:56       532121.acm.csv         532105 acm       coarse_mes       atorok       2010-04-13 16:21:00       1:27:12       532105 acm.csv         531876 acm       coarse_mes       atorok       2010-04-13 16:21:00       1:27:12       532105 acm.csv         53199.acm       STDIN       gshi       2010-04-07 10:31:35       530958 acm.csv         530958.acm       STDIN       gshi       2010-04-07 16:32:47       0:10:21       530954.acm.csv         530767.acm       coarse_mes       atorok       2010-04-07 14:41:13       1:29:45       530767.acm.csv         530610 acm       coarse_mes       atorok       2010-04-06 23:40:12       1:27:1       530610 acm.csv         530558.acm       coarse_mes       atorok       2010-04-06 23:95:5       1:28:56       530558 acm.csv         530562.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:7       530561 acm.csv         530518.acm       coarse_mes       atorok	<u>532171.acm</u>	coarse_mes	atorok	2010-04-13 22:14:47	1:27:30	532171.acm.csv		
532137.acm       coarse_mes       atorok       2010-04-13 19:15:36       1:28:54       532137.acm.csv         532121.acm       coarse_mes       atorok       2010-04-13 17:49:22       1:24:56       532121.acm.csv         532105.acm       coarse_mes       atorok       2010-04-13 16:21:00       1:27:12       532105.acm.csv         531876.acm       coarse_mes       atorok       2010-04-12 18:38:50       2:0:0       531876.acm.csv         531199.acm       STDIN       gshi       2010-04-07 09:93:045       33:54:39       531199.acm.csv         530958.acm       STDIN       gshi       2010-04-07 16:58:07       40:31:35       530958.acm.csv         530767.acm       coarse_mes       atorok       2010-04-07 16:32:47       0:10:21       530954.acm.csv         530610.acm       coarse_mes       atorok       2010-04-07 14:44:13       1:29:45       530767.acm.csv         530561.acm       coarse_mes       atorok       2010-04-06 23:40:12       1:27:1       530610.acm.csv         530562.acm       coarse_mes       atorok       2010-04-06 20:39:54       1:28:37       530562.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:45       530540.acm.csv         530518.acm       coarse_mes </td <td>532155.acm</td> <td>coarse_mes</td> <td>atorok</td> <td>2010-04-13 20:45:36</td> <td>1:27:58</td> <td>532155.acm.csv</td> <td></td> <td></td>	532155.acm	coarse_mes	atorok	2010-04-13 20:45:36	1:27:58	532155.acm.csv		
532121.acm       coarse_mes       atorok       2010-04-13       17:49:22       1:24:56       532121.acm.csv         532105.acm       coarse_mes       atorok       2010-04-13       16:21:00       1:27:12       532105.acm.csv         531199.acm       coarse_mes       atorok       2010-04-12       18:38:50       2:00       531876.acm.csv         531199.acm       STDIN       gshi       2010-04-09       09:30:45       33:54:39       531199.acm.csv         530958.acm       STDIN       gshi       2010-04-07       16:58:07       40:31:35       530958.acm.csv         530954.acm       STDIN       gshi       2010-04-07       16:32:47       0:10:21       530954.acm.csv         530054.acm       coarse_mes       atorok       2010-04-07       16:32:47       0:10:21       530954.acm.csv         530050.acm       coarse_mes       atorok       2010-04-07       14:41:13       1:29:45       530160.acm.csv         530555.acm       coarse_mes       atorok       2010-04-06       20:39:51       1:28:37       530562.acm.csv         530562.acm       coarse_mes       atorok       2010-04-06       1:28:37       530540.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06       1	<u>532137.acm</u>	coarse_mes	atorok	2010-04-13 19:15:36	1:28:54	532137.acm.csv		
532105.acm       coarse_mes       atorok       2010-04-13 16:21:00       1:27:12       532105.acm csv         531876.acm       coarse_mes       atorok       2010-04-12 18:38:50       2:00       531876.acm.csv         531199.acm       STDIN       gshi       2010-04-09 09:30:45       33:54:39       531199.acm.csv         530958.acm       STDIN       gshi       2010-04-07 16:58:07       40:31:35       530958.acm.csv         530954.acm       STDIN       gshi       2010-04-07 16:32:47       0:10:21       530954.acm.csv         5300610.acm       coarse_mes       atorok       2010-04-07 14:44:13       1:29:45       530767.acm.csv         530585.acm       coarse_mes       atorok       2010-04-07 14:44:13       1:29:45       530610.acm.csv         530610.acm       coarse_mes       atorok       2010-04-06 23:40:12       1:27:1       530610.acm.csv         530585.acm       coarse_mes       atorok       2010-04-06 20:39:54       1:28:37       530562.acm.csv         530540.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:45       530540.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06 16:11:24       1:28:45       530496.acm.csv         530496.acm       coarse_mes	532121.acm	coarse_mes	atorok	2010-04-13 17:49:22	1:24:56	532121.acm.csv		
531876.acm       coarse_mes       atorok       2010-04-12 18:38:50       2:0:0       531876.acm.csv         531199.acm       STDIN       gshi       2010-04-09 09:30:45       33:54:39       531199.acm.csv         530958.acm       STDIN       gshi       2010-04-07 16:58:07       40:31:35       530958.acm.csv         530954.acm       STDIN       gshi       2010-04-07 16:32:47       0:10:21       530954.acm.csv         530767.acm       coarse_mes       atorok       2010-04-07 14:44:13       1:29:45       530767.acm.csv         530610.acm       coarse_mes       atorok       2010-04-06 23:40:12       1:27:1       530610.acm.csv         530585.acm       coarse_mes       atorok       2010-04-06 22:09:55       1:28:36       530585.acm.csv         530562.acm       coarse_mes       atorok       2010-04-06 20:39:54       1:28:37       530562.acm.csv         530540.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:6       530540.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06 17:41:24       1:28:7       530518.acm.csv         530496.acm       coarse_mes       atorok       2010-04-06 16:11:24       1:28:45       530496.acm.csv         530496.acm       coarse_mes	532105.acm	coarse_mes	atorok	2010-04-13 16:21:00	1:27:12	532105.acm.csv		
531199.acm       STDIN       gshi       2010-04-09 09:30:45       33:54:39       531199.acm.csv         530958.acm       STDIN       gshi       2010-04-07 16:58:07       40:31:35       530958.acm.csv         530954.acm       STDIN       gshi       2010-04-07 16:32:47       0:10:21       530954.acm.csv         530767.acm       coarse_mes       atorok       2010-04-07 14:44:13       1:29:45       530767.acm.csv         530610.acm       coarse_mes       atorok       2010-04-06 23:40:12       1:27:1       530610.acm.csv         530585.acm       coarse_mes       atorok       2010-04-06 22:09:55       1:28:56       530585.acm.csv         530562.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:6       530540.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:6       530540.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06 17:41:24       1:28:6       530540.acm.csv         530496.acm       coarse_mes       atorok       2010-04-06 16:11:24       1:28:45       530496.acm.csv         530496.acm       coarse_mes       atorok       2010-04-06 16:11:24       1:28:45       530496.acm.csv         530496.acm       coarse_mes	531876.acm	coarse_mes	atorok	2010-04-12 18:38:50	2:0:0	531876.acm.csv		
530958.acm       STDIN       gshi       2010-04-07 16:58:07       40:31:35       530958.acm.csv         530954.acm       STDIN       gshi       2010-04-07 16:32:47       0:10:21       530954.acm.csv         530767.acm       coarse_mes       atorok       2010-04-07 14:44:13       1:29:45       530767.acm.csv         530610.acm       coarse_mes       atorok       2010-04-06 23:40:12       1:27:1       530610.acm.csv         530585.acm       coarse_mes       atorok       2010-04-06 22:09:55       1:28:56       530585.acm.csv         530562.acm       coarse_mes       atorok       2010-04-06 20:39:54       1:28:37       530562.acm.csv         530540.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:6       530540.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:7       530518.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06 16:11:24       1:28:7       530518.acm.csv         530496.acm       coarse_mes       atorok       2010-04-06 16:11:24       1:28:45       530496.acm.csv         530496.acm       coarse_mes       atorok       2010-04-06 14:42:19       1:27:53       530474.acm.csv	531199.acm	STDIN	gshi	2010-04-09 09:30:45	33:54:39	531199.acm.csv		
530954.acm       STDIN       gshi       2010-04-07 16:32:47       0:10:21       530954.acm.csv         530767.acm       coarse_mes       atorok       2010-04-07 14:44:13       1:29:45       530767.acm.csv         530610.acm       coarse_mes       atorok       2010-04-06 23:40:12       1:27:1       530610.acm.csv         530585.acm       coarse_mes       atorok       2010-04-06 22:09:55       1:28:56       530585.acm.csv         530562.acm       coarse_mes       atorok       2010-04-06 20:39:54       1:28:37       530562.acm.csv         530540.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:6       530540.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06 17:41:24       1:28:7       530518.acm.csv         530496.acm       coarse_mes       atorok       2010-04-06 16:11:24       1:28:7       530496.acm.csv         530474.acm       coarse_mes       atorok       2010-04-06 14:42:19       1:27:53       530474.acm.csv	530958.acm	STDIN	gshi	2010-04-07 16:58:07	40:31:35	530958.acm.csv		
530767.acm       coarse_mes       atorok       2010-04-07 14:44:13       1:29:45       530767.acm.csv         530610.acm       coarse_mes       atorok       2010-04-06 23:40:12       1:27:1       530610.acm.csv         530585.acm       coarse_mes       atorok       2010-04-06 22:09:55       1:28:56       530585.acm.csv         530562.acm       coarse_mes       atorok       2010-04-06 20:39:54       1:28:37       530562.acm.csv         530540.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:6       530540.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06 17:41:24       1:28:7       530518.acm.csv         530496.acm       coarse_mes       atorok       2010-04-06 16:11:24       1:28:45       530496.acm.csv         530474.acm       coarse_mes       atorok       2010-04-06 14:42:19       1:27:53       530474.acm.csv	530954.acm	STDIN	gshi	2010-04-07 16:32:47	0:10:21	530954.acm.csv		
530610.acm       coarse_mes       atorok       2010-04-06 23:40:12       1:27:1       530610.acm.csv         530585.acm       coarse_mes       atorok       2010-04-06 22:09:55       1:28:56       530585.acm.csv         530562.acm       coarse_mes       atorok       2010-04-06 20:39:54       1:28:37       530562.acm.csv         530540.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:6       530540.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06 17:41:24       1:28:7       530518.acm.csv         530496.acm       coarse_mes       atorok       2010-04-06 14:42:19       1:27:53       530474.acm.csv	530767.acm	coarse_mes	atorok	2010-04-07 14:44:13	1:29:45	530767.acm.csv		
530585.acm       coarse_mes       atorok       2010-04-06 22:09:55       1:28:56       530585.acm.csv         530562.acm       coarse_mes       atorok       2010-04-06 20:39:54       1:28:37       530562.acm.csv         530540.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:6       530540.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06 17:41:24       1:28:7       530518.acm.csv         530496.acm       coarse_mes       atorok       2010-04-06 16:11:24       1:28:45       530496.acm.csv         530474.acm       coarse_mes       atorok       2010-04-06 14:42:19       1:27:53       530474.acm.csv	530610.acm	coarse_mes	atorok	2010-04-06 23:40:12	1:27:1	530610.acm.csv		
530562.acm       coarse_mes       atorok       2010-04-06 20:39:54       1:28:37       530562.acm.csv         530540.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:6       530540.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06 17:41:24       1:28:7       530518.acm.csv         530496.acm       coarse_mes       atorok       2010-04-06 16:11:24       1:28:45       530496.acm.csv         530474.acm       coarse mes       atorok       2010-04-06 14:42:19       1:27:53       530474.acm.csv	530585.acm	coarse_mes	atorok	2010-04-06 22:09:55	1:28:56	530585.acm.csv		
530540.acm       coarse_mes       atorok       2010-04-06 19:10:42       1:28:6       530540.acm.csv         530518.acm       coarse_mes       atorok       2010-04-06 17:41:24       1:28:7       530518.acm.csv         530496.acm       coarse_mes       atorok       2010-04-06 16:11:24       1:28:45       530496.acm.csv         530474.acm       coarse_mes       atorok       2010-04-06 14:42:19       1:27:53       530474.acm.csv	530562.acm	coarse_mes	atorok	2010-04-06 20:39:54	1:28:37	530562.acm.csv		
530518.acm         coarse_mes         atorok         2010-04-06 17:41:24         1:28:7         530518.acm.csv           530496.acm         coarse_mes         atorok         2010-04-06 16:11:24         1:28:45         530496.acm.csv           530474.acm         coarse mes         atorok         2010-04-06 14:42:19         1:27:53         530474.acm.csv	530540.acm	coarse_mes	atorok	2010-04-06 19:10:42	1:28:6	530540.acm.csv		
530496.acm         coarse_mes         atorok         2010-04-06 16:11:24         1:28:45         530496.acm.csv           530474.acm         coarse mes         atorok         2010-04-06 14:42:19         1:27:53         530474.acm.csv	530518.acm	coarse_mes	atorok	2010-04-06 17:41:24	1:28:7	530518.acm.csv		
530474.acm coarse mes atorok 2010-04-06 14:42:19 1:27:53 530474.acm.csv	<u>530496.acm</u>	coarse_mes	atorok	2010-04-06 16:11:24	1:28:45	530496.acm.csv		
	530474.acm	coarse mes	atorok	2010-04-06 14:42:19	1:27:53	530474.acm.csv	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	-

**NCSA** 

#### **Power Profiling Walk Through**

#### AC Power Utilization



#### JSON Data

- Mouse-over value displays
- Under curve totals displayed
- If there is user interest, we may support calls to add custom tags from application



#### **AC GPU Cluster Power Considerations**

State	Host Peak	Tesla Peak	Host	Tesla power
	(Watt)	(Watt)	power factor	factor (pf)
			(pf)	
power off	4	10	.19	.31
start-up	310	182		
pre-GPU use idle	173	178	.98	.96
after NVIDIA driver module	173	178	.98	.96
unload/reload <sup>(1)</sup>				
after deviceQuery <sup>(2)</sup> (idle)	173	365	.99	.99
GPU memtest #10 (stress)	269	745	.99	.99
after memtest kill (idle)	172	367	.99	.99
after NVIDIA module	172	367	.99	.99
unload/reload <sup>(3)</sup> (idle)				
VMD Madd	268	598	.99	.99
NAMD GPU STMV	321	521	.97-1.0	.85-1.0 <sup>(4)</sup>
NAMD CPU only ApoA1	322	365	.99	.99
NAMD CPU only STMV	324	365	.99	.99

1. Kernel module unload/reload does not increase Tesla power

2. Any access to Tesla (e.g., deviceQuery) results in doubling power consumption after the application exits

3. Note that second kernel module unload/reload cycle does not return Tesla power to normal, only a complete reboot can

4. Power factor stays near one except while load transitions. Range varies with consumption swings



#### **Power Profiling Hardware (2)**

- Improved accuracy
- Increased granularity (every .20 seconds)
- Current flexible
- Voltage flexible
- 4 monitored ports
- ~\$50 total parts





#### **Application Power profile example**

- No GPU used
  - 240 Watt idle
  - 260 Watt computing on a single core

- Computing on GPU
  - 240 Watt idle
  - 320 Watt computing on a single core





#### **Improvement in Performance-per-watt**

 $e = p/p_a^*s$ 

p – power consumption of non-accelerated application  $p_a$  – power consumption of accelerated application s – achieved speedup

Application	t (sec)	t <sub>a</sub> (sec)	S	p (watt)	p <sub>a</sub> (watt)	е
NAMD	6.6	1.1	6	316	681	2.78
VMD	1,465.2	57.5	25.5	299	742	10.48
QMCPACK			61.5	314	853	22.6
MILC	77,324	3,881	19.9	225	555	8.1



#### **Speedup-to-Efficiency Correlation**

 The GPU consumes roughly double the CPU power, so a 3x GPU is require to break even





# **Applications**

- Applications
  - Cosmology
  - Computational chemistry
  - Quantum chromodynamics
  - Data mingin



#### Lattice QCD: MILC

- Simulation of the 4D SU(3) lattice gauge theory
- Solve the space-time 4D linear system  $M\phi = b$  using CG solver
  - $\phi_{i,x}$  and  $b_{i,x}$  are complex variables carrying a color index i = 1,2,3 and a 4D lattice coordinate x.
  - M = 2mal + D
  - *I* is the identity matrix
  - 2ma is constant, and
  - matrix *D* (called "Dslash" operator) is given by  $D_{x,i;y,j} = \sum_{\mu=1}^{4} \left( U_{x,\mu}^{F\,i,j} \delta_{y,x+\widehat{\mu}} - U_{x-\widehat{\mu},\mu}^{F\,\dagger\,i,j} \delta_{y,x-\widehat{\mu}} \right)$   $+ \sum_{\mu=1}^{4} \left( U_{x,\mu}^{L\,i,j} \delta_{y,x+3\widehat{\mu}} - U_{x-3\widehat{\mu},\mu}^{L\,\dagger\,i,j} \delta_{y,x-3\widehat{\mu}} \right)$

Input:	
D: dslash operator	
$A_h: D^+D + 4ma^2$ in high precision	
$A_1: D^+D + 4ma^2$ in low precision	
b: source vector	
x: guess solution vector	
Output:	
y: solution vector	
y ← 0	
i ← 0	
r ← b - A <sub>h</sub> x	D+K
d ← r	
δnew ← r <sup>T</sup> r	К1
$\delta_0 \leftarrow b^T b$	К1
while i < $i_{max}$ and $\delta > \delta_0 \varepsilon^2$ do	

q ← A <sub>1</sub> d	D
α ← δnew /(d <sup>I</sup> q)	K2
δ <sub>old</sub> ← δ <sub>new</sub>	
r ← r - ¤q	K3

$$\delta_{\text{new}} \leftarrow r^{\text{T}}r$$

if (high-precision update is needed)	
x ← x + αd	К4
у ← х + у	К5
r ← b - A <sub>h</sub> y	D+K7
δ <sub>new</sub> ← r <sup>T</sup> r	К1
x ← 0	
$\beta \leftarrow \delta_{new} / \delta_{old}$	
d ← r + βd	K8
else	
x ← x + αd	K6
d ← r + βd	
i ← i + 1	

Collaboration with Steven Gottlieb from U Indiana, Bloomington



#### **GPU Implementation strategy:** optimize for memory bandwidth

4D lattice



- flop-to-byte ratio of the Dslash operation is 1,146/1,560=0.73
- flop-to-byte ratio supported by the C2050 hardware is 1,030/144=7.5
- thus, the Dslash operation is memory bandwidth-bound

spinor data layout



link data layout





#### Parallelization strategy: split in T dimension

- 4D lattice is partitioned in the time dimension, each node computes T slices
- Three slices in both forward and backward directions are needed by the neighbors in order to compute new spinors



- dslash kernel is split into
  - interior kernel which computes the internal slices (2<t<T-3) of sublattice and the space contribution of the boundary sub-lattices, and
  - exterior kernel which computes the time dimenstion contribution for the boundary sub-lattice. The exterior kernel depends on the data from the neighbors.
- The interior kernel and the communication of boundary data can be overlapped using CUDA streams



#### **Results for CG solver alone**



one CPU node = 8 Intel Nehalem 2.4 Ghz CPU cores one GPU node = 1 CPU core + 1 C2050 GPU

lattice size 28<sup>3</sup>x96



#### **Results for entire application** (Quantum Electrodynamics)





#### References

#### GPU clusters

- V. Kindratenko, J. Enos, G. Shi, M. Showerman, G. Arnold, J. Stone, J. Phillips, W. Hwu, *GPU Clusters for High-Performance Computing*, in Proc. <u>IEEE International Conference on Cluster Computing</u>, <u>Workshop on Parallel</u> <u>Programming on Accelerator Clusters</u>, 2009.
- M. Showerman, J. Enos, A. Pant, V. Kindratenko, C. Steffen, R. Pennington, W. Hwu, QP: A Heterogeneous Multi-Accelerator Cluster, In Proc. <u>10th LCI International Conference on High-Performance Clustered Computing – LCI'09</u>, 2009.

#### Memory reliability

• G. Shi, J. Enos, M. Showerman, V. Kindratenko, *On testing GPU memory for hard and soft errors*, in Proc. <u>Symposium on Application Accelerators in High-Performance Computing – SAAHPC'09</u>, 2009

#### Power efficiency

• J. Enos, C. Steffen, J. Fullop, M. Showerman, G. Shi, K. Esler, V. Kindratenko, J. Stone, J. Phillips, *Quantifying the Impact of GPUs on Performance and Energy Efficiency in HPC Clusters*, In Proc. <u>Work in Progress in Green Computing</u>, 2010.

#### Applications

- S. Gottlieb, G. Shi, A. Torok, V. Kindratenko, *QUDA programming for staggered quarks*, In Proc. <u>The XXVIII International</u> <u>Symposium on Lattice Field Theory Lattice'10</u>, 2010.
- G. Shi, S. Gottlieb, A. Totok, V. Kindratenko, *Accelerating Quantum Chromodynamics Calculations with GPUs*, In Proc. <u>Symposium on Application Accelerators in High-Performance Computing</u> - SAAHPC'10, 2010.
- A. Titov, V. Kindratenko, I. Ufimtsev, T. Martinez, Generation of Kernels to Calculate Electron Repulsion Integrals of High Angular Momentum Functions on GPUs – Preliminary Results, In Proc. <u>Symposium on Application Accelerators in High-</u> <u>Performance Computing</u> - SAAHPC'10, 2010.
- G. Shi, I. Ufimtsev, V. Kindratenko, T. Martinez, *Direct Self-Consistent Field Computations on GPU Clusters*, In Proc. <u>IEEE</u> International Parallel and Distributed Processing Symposium – IPDPS, 2010.
- D. Roeh, V. Kindratenko, R. Brunner, *Accelerating Cosmological Data Analysis with Graphics Processors*, In Proc. <u>2nd</u> <u>Workshop on General-Purpose Computation on Graphics Processing Units – GPGPU-2</u>, pp. 1-8, 2009.

