500 Teraflops Heterogeneous Cluster
(Air Force largest interactive HPC)

Mr. Mark Barnell
HPC Director AFRL/RIT
Voice: (315) 330-3273
Email: Mark.Barnell@rl.af.mil
Introduction

• This will likely keep us in the world wide lead (certainly within DoD) for hosting the largest interactive supercomputer.

• The plan is to move the present Cell BE cluster to another facility.

• These machines are freely available to government researchers and our contractors.
What makes this advance possible?

• As the server market drove price-performance improvements that the HPC community leveraged over the past decade, now the gaming marketplace may deliver 10x-20x improvements (power as well).
  – $3800 3.2 GHz dual-quad core Xeon® , 96 Gflops (DP)- baseline system, Power 1000 Watts
  – $380 3.2 GHz PS3® with Cell Broadband Engine® 153 Gflops (SP), power 135 Watts
    • 1.6X Flops/board, 1/10th cost
  – $2000 Tesla C2050 (515Gflops (DP), 1.03Tflops (SP)), Power 225 Watts
    • 1/10th cost, 1/20th the power
PlayStation3 Fundamentals

- $380
- Cell BE ® processor
- 256 MB RDRAM (only)
- 160 GB hard drive
- Gigabit Ethernet (only)
- 153 Gflops Single Precision Peak
- 380 TFLOPS/$M
- Sony Hypervisor
- Fedora Core 7 or 9 Linux or YDL 6.2
- IBM CELL SDK 3.1

- 6 of 8 SPEs available
- 25.6 GB/sec to RDRAM
- ~110 Watts
AFRL/RIT Horus Cluster

10 - 1U Rack Servers

- **26 Tflops**
  - Supports TTCP efforts
  - 18 General Purpose Graphical Processor Units (GPGPUs) Cluster

NVIDIA C2050

1.1TFLOPS SP
515GFLOPS DP
Key Questions

• Which codes could scale given these constraints?
• Can a hybrid mixture of PS3s and traditional servers mitigate the weaknesses of the PS3s alone and still deliver outstanding price-performance?
• What level of effort is required to deliver a reasonable percentage of the enormous peak throughput?
• A case study approach is being taken to explore these questions
Early Access System Approach

• A 53 TeraFLOPS cluster of Playstation® 3s has been built at AFRL Information Directorate in Rome, NY to provide **early access** to the IBM Cell Broadband Engine® chip technology included in the low priced commodity gaming consoles.

• A heterogeneous cluster with powerful subcluster headnodes is used to balance the architecture in light of PS3 memory and input/output constraints
  – 14 subclusters each with 24 PS3s and a headnode

• Interactive usage (at least for now)

• Available to HPCMP community for experimentation
The Cell Cluster has a peak performance of 51.5 Teraflops from 336 PS3s and additional 1.4 TF from the headnodes on its 14 subclusters.

- Cost: $361K ($257K from HPCMP)
  - PS3s 37% of cost
- Price Performance: 147 TFLOPS/$M
- The 24 PS3s in aggregate contain 6 GB of memory and 960 GB of disk. The dual quad-core Xeon headnodes have 32 GB of DRAM and 4 TB of disk each.
500 TFLOPS Notional Architecture
(2010)

- ~300 Tflops from 2000 PS3s
- ~200 Tflops from GPGPUs on subcluster headnodes
- Cost: ~$2M
500 TFLOPS Architecture

CONDOR CLUSTER
Online: October - November 2010

- Approx. 270 TFLOPS from 1,760 PS3s
  - 153 GFLOPS/PS3
  - 80 subclusters of 22 PS3s
- Approx. 230 TFLOPS from subcluster headnodes
  - 2 GPGPU (2.1 TFLOPS / headnode)
  - 84 headnodes (Intel Nehalem 5660 dual socket Hexa (12 cores))
  - *Horus Cluster (~26 Tflops)
- Cost: Approx. $2M
- Total Power 300KW
CONDOR Node (Dual Nahlem x5650, 24 GB Ram, 2TB HD, 1200W PS, 2 Tesla GPGPUs, 40Gb/s Inf, Dual 10Gb (2.5 Tflops SP or 1.2 Tflops DP)
This project provides the HPCMP community with early access to HPC scale commodity multicore through a 336 node cluster of PS3 gaming consoles (53 TF).

Applications leveraging the >10X price-performance advantage include:

- Large scale simulations of neuromorphic computing models
- GOTCHA radar video SAR for wide area persistent surveillance
- Real-time PCID image enhancement for space situational awareness

Neuromorphic example: Robust recognition of occluded text
The driving application behind developing a 53 TF class cluster was to support basic research into alternative neuromorphic computing architectures.

The first of these to be optimized for the PS3 was the “Brain-State-In –A-Box” (BSB)—looking for 1M BSBs simulating in real time.

Optimized the BSB for the PS3 and achieved 18 GFLOPS on each core of the PS3 [6]. Across the 6 cores, 108 GFLOPS/PS3, over 70% of peak was sustained.

12 staff week effort for first PS3 optimization experience

Constructing hybrid simulations with BSBs and “Confabulation” models
Minicolumn Model
Hybrid: Attractor + Geometric Receptors

Mechanisms identified during initial effort are being applied to a closely neuromorphic columnar model we are emulating on a Cell-BE Cluster.

Literature reviews: minicolumn anatomy, cortical anatomy, cortical modeling, Cog Psyc, Neural Sci.

Explored attractors (BSB, Willshaw, PINN, Sparse Distributed Memory, Limit cycle) & arrays of (Erzatz Brain, Liquid State Machines).

Assessment of Confabulation: algorithm complexity, efficacy, acceleration.

Development of Hybrid model:
Simple/Complex cell minicolumn, functional columns, full scale V1.

Spiky Neuron Dynamical Modeling: emulation exercise – 64 minicolumns assembled as a functional column.
Neuromorphic Vision System

SONY EVI-HD1 Cameras
- 1080p @ 29.97 fps (60Hz)
- 10x Optical / 4x Digital Zoom

JBCC
- Pointing (FOV Modeling)
- Stereo Vision Disparity (MAE)
- Slice Storage
- Object Recognition (MAE Threshold)

196 PS3s using BSB models to compute:
- Orientation lines
- Color
- Light intensity

100K JBI Pub/Sub
- Publishing 60 fps
- 3,236 msgs/frame
- Payload: 80 MB/frame
- 37.5 gigabits/sec

Sensor to HPC
Mapping V1 Model to HPC
One Subfield per PS3 Node

- Full V1 (196 Subfields)
- A “neighborhood” of 6 subfields.
- 196 PS3 nodes emulating a Full V1

Connectivity pattern of a Functional Column
Intense local communications within 3 mm

1 Subfield:
- 2 Field of Views
- 64 Functional Columns per FOV

1 Functional Column:
- 64 minicolumns

1 Minicolumn:
- 32 Element BSB
- “Simple & complex cells”
- “Readout” cells

Consensus of percepts

196 Subfields x 2 FOV x 64 FC’s x 64 MC’s = 1,605,632 Minicolumns
Results: Neuromorphic Modelling

- Cell BE Cluster networking infrastructure is more than up to the challenge of handling the I/O from most I/O intensive models under examination (400-500 Hz update far exceeds 100 Hz real-time need)
Hybrid Cognitive Model for Text Recognition

...but beginning to perceive that the handcuffs were not for me and that the military had so far got....

Perception based on neural network models

Prediction

Word Level Confabulation

...but beginning to perceive that the handcuffs were not for me and that the military had so far got....

Sentence Level Confabulation

Prediction
Confabulation Architecture
Core & Processor Level Parallelism

Core level parallelism
Multi-threading, Shared memory

Processor level parallelism
Loosely synchronous, MPI communication

2009

Now

派遣 Images
接收 Letters

BSB BSB BSB BSB

Word Level Confab.

Sentence Confab.

8-core Xeon

24 PS3s

12-core Intel Xeon
Processor

Performance Monitor

22 PS3s
## Performance Evaluation

### Computing power from Cell processors (GFLOPS)

<table>
<thead>
<tr>
<th></th>
<th>PS3 node 1 Cell Processor</th>
<th>Sub-cluster 1 head-node + 24 PS3</th>
<th>HPC cluster 14 sub-clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computing power from Cell processors (GFLOPS)</td>
<td>75</td>
<td>1800</td>
<td>25200</td>
</tr>
<tr>
<td>Character recognition peak performance (characters / sec)</td>
<td>48</td>
<td>1152</td>
<td>16128</td>
</tr>
<tr>
<td>Word confabulation peak performance (words / sec)</td>
<td>N/A</td>
<td>30</td>
<td>420</td>
</tr>
<tr>
<td>Sentence confabulation peak performance (sentences / sec)</td>
<td>N/A</td>
<td>160</td>
<td>2240</td>
</tr>
<tr>
<td>Overall typical text recognition performance (sentences / sec)</td>
<td>N/A</td>
<td>4.3</td>
<td>59.9</td>
</tr>
</tbody>
</table>
Video Synthetic Aperture Radar
Backprojection Case

• This algorithm is expensive computationally, but allows SAR radar images to focus each pixel independently, accounting for variations in elevation.

• This algorithm was accelerated >300X over original XEON code and achieved 40% of peak (60 GFLOPS sustained) on each PS3.

• 8 PS3s and headnode flown in 2007 for 1.5km spot

• 96 PS3s demonstrated 5km spot processing in Lab in May 08

• 9 1U Servers (8 with dual GPGPUs)
  – 2km x10km swath, 2.2 Tflops sustained

• 20 KM spot-72 Tflops, 40 KM spot 350 Tflops
Results: Gotcha VideoSAR Scalability

- At 256 PS3s, each send 6 MB/sec and receives 8.4 MB/sec while headnodes each receive 200 MB/sec and send 140 MB/sec
SAR Image Formation using GPGPUs

9 KM SPOT
500Hz PRF
16 Tesla Boards

64 MB Pulse Packs (500 pulse segments)

Master Nodes: Pulse Compression, Form, Store Image, publish frames

Node 9: Collects aux data, merges radar signals and sends to Master Nodes

Each Node does Pulse compression

Node 9 merges segments and sends them via TCP

Master Nodes:  Pulse Compression, Form, Store Image, publish frames

Node 9 merges segments and sends them via TCP
High Definition (HD) Video Processing Case

- This case study employed 3 PS3s and a headnode to process 1080p grayscale HD video (52 MB/sec at 1080x1920) in real time.

- The sum of absolute differences algorithm was first optimized to process 64 frames/second at HD resolution (1920x1080). 21 GFLOPS was sustained on one PS3.

- Flux tensor optimized to 60 GFLOPS (40% of peak), 92X Xeon, in 3 staff months.

- The headnode played a key role of both archiving and disseminating the HD video to the PS3s in parallel streams.
Large Matrix-Matrix Multiply Case

• Matrix multiplication runs near peak for small matrices on a single PS3—but can it scale across a cluster and still have excellent performance?

• In theory, yes!
  – Source portions of the A matrix from local disk
  – Multicast the B matrix in on gigabit ethernet

• Progress to date:
  – Extended Dresden square MM to rectangular MM
  – Tested Ethernet (~90 MB/s) and Disk (35 MB/s) capabilities
  – Combining the pieces (where theory meets practice)
Large Matrix-Matrix Multiply Case

I/O Rates for Peak MM Performance Given Working Memory Available

Potential region for peak performance

Gigabit Ethernet input rate (MB/sec)

Necessary Disk Read rate (MB/sec)
Conclusions

• The large computer game marketplace is bringing forward products with tremendous computational performance at commodity prices.

• By November, 2010 a world-class 500 TFLOPS interactive supercomputer should come online at RRS

• While not all codes will be able to benefit, given memory and I/O constraints, the set is enlarged by complementing the PS3s with powerful headnodes in a subcluster configuration with dual GPGPUs (C1060s & C2050s)

• Several applications scaling very well and achieving significant percentage of Cell BE peak performance

• SAR backprojection algorithm scaling well on GPGPUs
Questions/Comments?