Roadrunner: 
A Fast But Unusual Bird, 
and Supercomputer

HPC User Forum
April 20, 2009

Ken Koch
Roadrunner Technical Manager,
Computer, Computational, and Statistical Sciences Division,
Los Alamos National Laboratory

Work presented was performed by a large team of Roadrunner project staff!
The messages this talk will convey are:

• Why Roadrunner? Why Cell?
  • A bold but important step toward the future

• What does Roadrunner look like?
  • Cluster-of-clusters with node-attached Cell blades

• Concepts for Programming Roadrunner
  • MPI, Opteron+Cell, “local-store” memory & DMA transfers

• Status and plans for Roadrunner
  • Timeline
  • Applications to date
A Roadrunner is born

using Cell processors as accelerators
Microprocessor trends have changed

• Moore’s law still holds, but is now being realized differently
  • More cores per chip and not all cores need be the same
  • Decreased memory bandwidth and capacity per core
  • Key findings of Jan. 2007 IDC Study: “Next Phase in HPC”
    • new ways of dealing with parallelism will be required
    • must focus more heavily on bandwidth (flow of data) and less on processor

From Burton Smith, LASC-06 keynote, with permission
The Cell processor is an (8+1)-way heterogeneous parallel processor

- Cell Broadband Engine (CBE*) developed by Sony-Toshiba-IBM
  - used in Sony PlayStation 3
- **8** Synergistic Processing Elements (SPEs)
  - 128-bit vector cores
  - 256 kB local memory (LS = Local Store)
  - Direct Memory Access (DMA) engine (25.6 GB/s each)
  - Chip interconnect (EIB)
  - Run SPE-code as POSIX threads (SPMD, MPMD, streaming)
- 1 PowerPC PPE runs Linux OS

**Original** Cell performance:
- 204.8 GF/s SP & 13.65 GF/s DP
- 512 MB @ 25.6 GB/s XDR memory
- Insufficient for a Petaflop/s machine for physics simulations

* trademark of Sony Computer Entertainment, Inc.
IBM created PowerXCell 8i

**Performance Enhancements/Scaling Path**

- **Cell BE (1+8)**
  - 90nm SOI
  - 2006

- **Cell/B.E. (1+8)**
  - 65nm SOI
  - 2007

- **Continued shrinks**
  - 2008

- **Cell/B.E. (1+8)**
  - 45nm SOI
  - 2009

- **Next Gen (2PPE'+32SPE')**
  - 45nm SOI
  - ~1 TF-SP (est.)

**PowerXCell 8i chip:**
- Used in Roadrunner
- 102.4 GF/s double precision
- 4 GB DDR2 @ 25.6 GB/s

**Cost Reduction Path**

- **PowerXCell is IBM’s name for this new enhanced double-precision (eDP) Cell processor variant**

All future dates and specifications are estimations only; Subject to change without notice. Dashed outlines indicate concept designs.
Los Alamos has a history in hybrid & petascale computing efforts led to Roadrunner taking a different path to a petascale system.
Roadrunner was delivered to LANL in Summer 2008

- New Hybrid Computing Model & Software
- New Triblade Hybrid Compute Node
- IBM Global Engineering Solutions was responsible for Development, Manufacturing and Support
IBM built hybrid nodes in Rochester, MN and assembled the system in Poughkeepsie, NY
Fully Assembled Roadrunner
Roadrunner broke the 1 Petaflop/s mark on May 26th, 2008

Matrix: ~5 trillion entries
Calculation: ~2 hours!

Performance: 1.026 Petaflop/s

Only 4 days after the full machine was finally assembled!
Roadrunner is a TOP performer!

<table>
<thead>
<tr>
<th>#</th>
<th>SITE</th>
<th>SYSTEM</th>
<th>Cores</th>
<th>TF/sec</th>
<th>MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DOE/NNSA/LANL United States</td>
<td>Roadrunner, QS22/LS21/IB</td>
<td>129600*</td>
<td>1105</td>
<td>2.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PowerXCell 8i, IBM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>DOE/ORNL United States</td>
<td>Jaguar, XT5,</td>
<td>150152</td>
<td>1059</td>
<td>6.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Opteron-QC,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cray</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>NASA Ames Research Center</td>
<td>Pleiades, Altix ICE &amp; IB,</td>
<td>51200</td>
<td>487</td>
<td>2.09</td>
</tr>
<tr>
<td></td>
<td>United States</td>
<td>Xeon-QC, SGI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>DOE/NNSA/LLNL United States</td>
<td>BGL, Blue Gene/L, PowerPC, IBM</td>
<td>212992</td>
<td>478</td>
<td>2.33</td>
</tr>
<tr>
<td>5</td>
<td>Argonne National Laboratory</td>
<td>Intrepid, Blue Gene/P, PowerPC,</td>
<td>163840</td>
<td>450</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td>United States</td>
<td>IBM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Texas Adv. Comp. Center</td>
<td>Ranger, SunBlade &amp; IB</td>
<td>62976</td>
<td>433</td>
<td>2.00</td>
</tr>
<tr>
<td></td>
<td>United States</td>
<td>Opteron-QC, Sun</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Roadrunner core count includes 8 SPEs in each Cell; Opteron+Cell-PPE cores is only 25920

#1 on the TOP500 (Nov. 2008)

#7 on the Green500 (Nov. 2008)

#5 & #6 are Roadrunner single CUs at LANL & IBM
Roadrunner System Configuration

See the LANL Roadrunner web site at end for more details
Roadrunner Phase 3 is Cell-accelerated, not a cluster of Cells

Node-attached Cells is what makes Roadrunner different!
A Roadrunner TriBlade node integrates Cell and Opteron blades

New PowerXCell 8i chips & QS22 blade

Two QS22’s with 2 Cells each

New part

Expansion blade

LS21 with two dual-core Opterons
A Roadrunner TriBlade node integrates Cell and Opteron blades

- **QS22** is an IBM Cell blade containing two new enhanced double-precision (eDP/PowerXCell™) Cell chips
- Expansion blade connects two QS22 via four PCI-e x8 links to LS21 & provides the node’s Mellanox ConnectX IB 4X DDR cluster attachment
- **LS21** is an IBM dual-socket Opteron blade
- 4-wide IBM BladeCenter packaging
- Roadrunner Triblades are completely diskless and run from RAM disks with NFS & Panasas only to the LS21
- Node design points:
  - One Cell chip per Opteron core
  - ~400 GF/s double-precision & ~800 GF/s single-precision
  - 16 GB Opteron memory PLUS 16 GB Cell memory
  - 1 PCI-E x8 to each Cell
A Connected Unit (CU) forms a building block

BC-H chassis 1

TriBlade 1
TriBlade 2
TriBlade 3

Voltaire IB4x DDR Switch

BC-H chassis 60

TriBlade 178
TriBlade 179
TriBlade 180

To 2nd Stage Switches

IB 4x DDR 2+2 GB/s
10 GigE 1+1 GB/s

10 GigE to file systems & LANs

2U I/O Node 1

2U I/O Node 12

2U Service Node
A Connected Unit (CU) is a powerful cluster

**Connected Unit Specifications:**

- 360 1.8 GHz dual-core Opterons
  - 2.59 TF DP peak Opteron
  - 2.88 TB Opteron memory
  - 24 2.6 GHz dual-core Opterons in I/O nodes

- 720 PowerXCell chips
  - 73.7 TF DP peak Cell
  - 2.88 TB Cell memory
  - 18.4 TB/s Cell memory BW

- 192 IB 4X DDR cluster links
  - 768 GB/s aggregate BW (bi-dir)
  - 384 GB/s bi-section BW (bi-dir)
  - 24 10 GigE I/O links on 12 I/O nodes
  - 24 GB/s aggregate I/O BW (uni-dir) (IB limited)

- 180 TriBlade
  - (1 LS21 + 2 QS22)
  - compute nodes

- 12 IBM x3655 I/O nodes
  - (dual 10 GigE each)
  - (dual-socket dual-core)

- 192 cluster nodes
  - 96 2nd-stage links

Voltaire 288-port IB 4x DDR
Now build a cluster-of-clusters...

Eight Voltaire IB 4x DDR 2<sup>nd</sup> Stage Switches

2<sup>nd</sup>-stage switches form a half-bandwidth fat-tree

17 CUs with CU switches, 3264 IB nodes

Extra 2<sup>nd</sup>-stage switch ports allow expansion up to 24 CUs

Operated by the Los Alamos National Security, LLC for the DOE/NNSA
Roadrunner is a petascale system in 2008

- 3,060 Compute Nodes
- 6,120 dual-core Opterons
- 44.1 TF DP peak Opteron
- 49 TB Opteron memory
- 204 I/O nodes w/ 408 Opterons

- 12,240 PowerXCell 8i chips
- 1.33 PF DP peak Cell
- 2.59 PF SP peak Cell
- 49 TB Cell memory
- 313 TB/s Cell memory BW

- 2-stage IB 4X DDR interconnect
  - 13.1 TB/s aggregate BW (bi-dir) (1st stage)
  - 6.5 TB/s aggregate BW (bi-dir) (2nd stage)
  - 3.3 TB/s bi-section BW (bi-dir) (2nd stage)
  - 204 I/O nodes with 408 10 GigE links to a Panasas parallel file system

17 CU clusters
3,060 compute nodes
12 links per CU to each of 8 switches
Eight 2nd-stage IB 4X DDR switches
Roadrunner at a glance

- Cluster of 17 Connected Units (CU)
  - 12,240 IBM PowerXCell 8i chips
  - 1.33 Petaflop/s DP peak (Cell)
  - 1.026 PF sustained Linpack (DP)
  - 6,120 (+408) AMD dual-core Opterons
  - 44.1 (+4.4) Teraflop/s peak (Opteron)

- Fedora Linux
  - On LS21 & QS22 blades & I/O & service nodes

- SDK for Multicore Acceleration
  - Cell compilers, libraries, tools

- xCAT Cluster Management
  - System-wide GigEnet network

- InfiniBand 4x DDR fabric
  - 3264 total nodes; 2-stage fat-tree; all-optical cables
  - Full bi-section BW within each CU
    - 384 GB/s (bi-directional)
  - Half bi-section BW among CUs
    - 3.26 TB/s (bi-directional)

- ~100 TB aggregate memory
  - 49 TB Opteron (compute nodes)
  - 49 TB Cell

- ~200 GB/s sustained File System I/O:
  - 204x2 10GE Ethernets to Panasas

- 2.35 MW Power:
  - 0.437 GF/Watt

- Area:
  - 280 racks
  - 5200 ft²
Programming Concepts
Programming Approaches for Roadrunner

Host Centric view

Opteron Process
Main Memory Footprint
Cell Process
Work Blocks

Accelerator Centric view

Message relay

Function offload

Opteron Process
Main Memory Footprint
Cell Process
Work Blocks

Cell Process
Main Memory Footprint
Cell Process
Work Blocks

Message relay

SPE view
DMA & Local Store
Multi-Buffering
SIMD vector

Opteron Process
Message Relay
Cell Process
Work Blocks

Opteron Process
Message Relay
Cell Process
Work Blocks

PowerPC
DMA & Local Store
Multi-Buffering
SIMD vector

Los Alamos National Laboratory
Operated by the Los Alamos National Security, LLC for the DOE/NNSA
Three types of processors work together

- **Parallel computing on Cell**
  - *data partitioning & work queue pipelining*
  - *process management & synchronization*

- **Remote communication to/from Cell**
  - *data communication & synchronization*
  - *process management & synchronization*
  - *computationally-intense offload*

- **MPI remains as the foundation**
Roadrunner nodes have a memory hierarchy

- **QS22 Cell blades**:
  - 256 KB of “working” memory (per SPE)
  - 25.6 GB/s off-SPE BW

- **Opteron blade**:
  - 8 GB of NUMA shared memory (per node)
  - 16 GB of distributed memory (per node)

- **Cell chip** per Opteron core:
  - 4 GB of memory (per core)
  - 8 GB of shared memory (per socket)
  - 5.4 GB/s/core

- **PCIe x8 (2 per blade)** (2 GB/s, 2 us)

- **ConnectX IB 4X DDR** (2 GB/s, 2 us)

- **“equal memory size” concept**
How do you keep the 256KB SPEs busy?

Break the work into a stream of pieces

problem domain of a Cell processor

grid tiles or particle bundles (can include ghost zones)

data chunks stream in & out of 8 SPEs using async DMAs and triple-buffering
Put it all together: MPI+DaCS+DMA+SIMD

- DMAs are simply block memory transfers
  - HW asynchronous (no SPE stalls)
  - DDR2 memory latency and BW performance

Compute & memory DMA transfers are overlapped in HW!

MPI & DaCS can also be fully asynchronous

- DMA Get (first prefetch)
  - Switch work buffers
- DMA Get (prefetch)
- DMA Wait (complet current)
- Compute
- DMA Put (store behind)
- DMA Wait (previous put)
  - Switch work buffers
- DMA Wait (put)
IBM-ALF is a simple work-queue approach for abstracting parallelism directly to SPEs

**ALF** = Accelerated Library Framework
Programming approach has now been demonstrated and is Tractable

- Two levels of parallelism:
  - *node-to-node*: MPI & *DaCS-MPI-DaCS relay*
  - *within-Cell*: threads, pipelined DMAs, & SIMD

- Large-grain computationally intense portions of code are split off for Cell acceleration within a node process
  - *Usually an entire tree of subroutines*
  - *This is equivalent to “function offload” of entire large algorithms*

- Threaded fine-grained parallelism introduced within the Cell itself
  - *Create many-way parallel pipelined work units for the 8 SPEs*
  - *Good for both multicore/manycore chips and heterogeneous chip trends with dwindling memory bandwidth*

- Communications during Cell computation are possible between Cells via DaCS-MPI-DaCS relay approach

- Considerable flexibility and opportunities exist beyond this approach
Five Waves of Roadrunner Applications Codes
Recent Roadrunner History

- At IBM POK
- Delivery
- Accept
- System & code stabilization
  - Open Science projects
  - Connect File Systems
  - Institutional Computing projects
  - 2 additional Open CUs: Cerrillos system

- 1st CU built
- 18 CUs built
- 2nd CU built
- 1.026 PF TOP500 on 17 CUs
- 1.105 PF TOP500 on 18 CUs

Operated by the Los Alamos National Security, LLC for the DOE/NNSA
Five Waves of Roadrunner Application Codes

   - Proof of Cell & Hybrid programming capability: 4 codes
   - Prototype hardware: old Cell/QS20 blades & very first eDP Cell

2. Full-System Pre-Acceptance Testing (June – Nov. 2008)
   - Gordon Bell finalists: VPIC & SPaSM
   - PetaVision (sustained 1+ single-precision-PF!)
   - PPM (Paul Woodward)

   - 8 codes & 10 projects today

4. Institutional Computing (starting May 2009 on Cerrillos – 2 CUs)
   - 19 new projects to start

5. Classified ASC Use (starting Oct. 2009)
Roadrunner Open Science and Institutional Computing

- HIV genealogy for drug design
- atomistic bioenergy
- more VPIC & SPaSM
- astrophysics

- nanowire formation
- reacting-compressible DNS turbulence
- ocean and climate modeling
- wildfires
- seismic waves
- discrete event simulation
- piecewise-parabolic method (PPM)
Exciting opportunities among the 10 selected proposals for Roadrunner Open Science

<table>
<thead>
<tr>
<th>Proposal</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinetic Thermonuclear Burn Studies with VPIC on Roadrunner</td>
<td>VPIC</td>
</tr>
<tr>
<td>Multibillion-Atom Molecular Dynamics Simulations of Ejecta Production and Transport using Roadrunner</td>
<td>SPaSM</td>
</tr>
<tr>
<td>New frontiers in viral phylogenetics</td>
<td>ML</td>
</tr>
<tr>
<td>Three-Dimensional Dynamics of Magnetic Reconnection in Space and Laboratory Plasmas</td>
<td>VPIC</td>
</tr>
<tr>
<td>The Roadrunner Universe</td>
<td>MC³</td>
</tr>
<tr>
<td>Implicit Monte Carlo Calculations of Supernova Light-Curves</td>
<td>IMC + Rage</td>
</tr>
<tr>
<td>Instabilities-Driven Reacting Compressible Turbulence</td>
<td>CFDNS</td>
</tr>
<tr>
<td>Cellulosomes in Action: Peta-Scale Atomistic Bioenergy Simulations</td>
<td>GROMACS</td>
</tr>
<tr>
<td>Parallel- replica dynamics study of tip-surface and tip-tip interactions in atomic force microscopy and the formation and mechanical properties of metallic nanowires</td>
<td>PAR-REP + CellMD</td>
</tr>
<tr>
<td>Saturation of Backward Stimulated Scattering of Laser In The Collisional Regime</td>
<td>VPIC</td>
</tr>
</tbody>
</table>

Indicates new work

Indicates new + old
The LANL Roadrunner web site is

http://www.lanl.gov/roadrunner/

Roadrunner architecture
Early applications efforts
Upcoming Open Science efforts
Cell & hybrid programming
Computing trends
Related Internet links