Roadrunner and hybrid computing

Ken Koch & Andy White
Roadrunner Project
Los Alamos National Laboratory

Salishan Conference on High-Speed Computing

April 26, 2007
Outline

1. Cell & Hybrid
2. Roadrunner Architecture (original & improved)
3. Programming & Applications
Cell & Hybrid
Cell Chip

- **Cell Broadband Engine™** * (Cell BE)
  - Developed under Sony-Toshiba-IBM efforts
  - Current Cell chip is used in the Sony PlayStation 3
- An 8-way heterogeneous parallel engine

Each of the 8 SPEs are 128 bit (e.g. 2-way DP-FP) vector engines w/ 256KB of Local Store (LS) memory & a DMA engine.

They can operate together or independently (SPMD or MPMD).

- ~200 GF/s single precision
- ~15 GF/s double precision (current chip)

* Trademark of Sony Computer Entertainment, Inc.
Cell Broadband Engine

Heterogeneous: 1PPU + 8 SPUs
Cell Broadband Engine Architecture™
Technology Competitive Roadmap

Performance Enhancements/Scaling

Cost Reduction

Cell BE (1+8) 90nm SOI

Continued shrink

Enhanced Cell BE (1+8eDP) 65nm SOI

1TF Processor 45nm SOI

Cell HPC chip:
To be used in Roadrunner
~100 GF/s double precision

2006
2007
2008
2009
2010

All future dates are estimations only; Subject to change without notice.
Cell Broadband Engine™ Blade
The first in a line of planned offerings using Cell Broadband Engine technology

Performance

2006

Cell BE-Based Blade
QS20
2 Cell BE Processors
Single Precision Floating Pt Affinity
1 GB Memory
Up to 4X PCI Express™

Available: Now

SDK 1.1

GC: 2H06

Advanced Cell BE-Based Blade
2 Cell BE Processors
Single Precision Floating Pt Affinity
2 GB Memory
Up to 16X PCI Express

Target Availability: 2H07

SDK 2.0

Target Availability: Dec06

SDK 2.1

Target Availability: 1H07

2007

Enhanced Cell BE-based Blade
2 Enhanced Cell BE Processors
SP & DP Floating Point Affinity
Up to 16 GB Memory (RR:8GB)
Up to 16X PCI Express

Target Availability: 1H08

SDK 3.0

2008

All future dates are estimations only; Subject to change without notice.

Los Alamos National Laboratory
Established 1943
Operated by the Los Alamos National Security, LLC for the DOE/NNSA
Salishan Roadrunner 7
The times they are a changin

Bob Dylan (1964)
Microprocessor trends are changing

- Moore’s law still holds but is now being realized differently
  - Clock frequency, chip power, & instruction-level-parallelism (ILP) have all plateaued
  - Multi-core is here today and manycore (≥ 32) looks to be the future
  - Complexity of shared memory and cache coherency for multi-core designs is likely not scalable to manycore designs
  - Memory bandwidth and memory capacity per core are headed downward (predominantly caused by increased core counts)
  - 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

- References:
  - IDC report #205025, January 2007
  - UC Berkeley UCB/EECS-2006-183
  - LASCI-06 Burton Smith keynote “Reinventing Computing”
Accelerators are outperforming CPUs
Before Roadrunner …

- Deep Blue for chess (IBM SP-2: 30 RS6K + 480 chess chips) (circa 1997)
- Grape-6 for stellar dynamics w/ custom chips) (circa 2000-2004)
  - [http://grape.astron.s.u-tokyo.ac.jp/~makino/grape6.html](http://grape.astron.s.u-tokyo.ac.jp/~makino/grape6.html)
- Various FPGA supercomputers from system vendors:
  - SRC-6 (w/ MAP)
  - Cray XD1 (w/ Application Acceleration)
  - SGI Altix (w/ RASC)
- Titech TSUBAME (w/ some Clearspeed) (2006)
  - [http://www.gsic.titech.ac.jp/English/Publication/pressrelease.html.en](http://www.gsic.titech.ac.jp/English/Publication/pressrelease.html.en)
- RIKEN MDGrape-3 “Protein Explorer” (w/ custom chips) (2006)
- Terra Soft’s Cell E.coli/Amoeba PS3 Cluster (cluster of 1U PlayStation 3 development systems) (2007)
  - [http://www.hpcwire.com/hpc/967146.html](http://www.hpcwire.com/hpc/967146.html)
Hybrid computing is here to stay

- Highly multi-core: 8, 16, 32, ...
  - IBM Cell, AMD Fusion, Intel Polaris, NVidia G8800
  - Distributed memory at core level

- Co-processors & accelerators
  - FPGAs, GPGPU, Clearspeed CSX600, IBM Cell, XtremeData XD1000, Nvidia G80, AMD Stream Processor

- Connection standards
  - AMD Torrenza, Intel/IBM Geneseo, AMD HyperTransport Initiative

- Programming
  - IBM Roadrunner ALF & DaCS libraries, RapidMind, Peakstream, Impulse C, Stanford’s Sequoia, NVidia CUDA, Clearspeed C, Mercury MFC, stream programming

- Heterogeneous architectures
  - within processor itself (e.g. Cell)
  - at the board level (e.g. AMD’s Torrenza)
  - on the same bus (e.g. CPU+GPUs, Intel’s Geneseo)
  - within a node (e.g. Roadrunner)
  - within a cluster
Roadrunner
Roadrunner project is a partnership with IBM

- Contract contract signed September 8, 2006 with IBM.

- Critical component of stockpile stewardship
  - **Phase 1** (Base system) supports near-term mission deliverables
  - **Phase 2** (Cell up-grade) supports pre-Final assessment
  - **Phase 3** (Hybrid final system)
    - Achieves PetaFlops level of performance
    - Demonstrates new paradigm for high performance computing

- Accelerated vision of the future
  - New programming paradigm
  - Faster processors
  - Still leveraging the marketplace

Cell processor (2007, 100 GF)

100x in 14 yrs
8 vector units each

CM-5 board (1994, 1 GF)
Roadrunner Hierarchy (Phase 1 Base)

Multiple Cluster Base System

14 CU clusters

2nd stage InfiniBand interconnect (multiple switches)
Roadrunner Hierarchy (Phase 3)

Final System with Cell Blade Accelerators

15 CU clusters

2nd stage InfiniBand interconnect (multiple switches)
Planned Roadrunner Phase 3 (Final System)

“Connected Unit” cluster
144 quad-socket
dual-core nodes
(138 w/ 4 dual-Cell blades
connected w/ 4 IB 4X SDR links)

8,640 dual-core Opterons
• 76 TeraFlop/s (total)

16,560 eDP Cell chips
• 1.7 PetaFlop/s (Cell)

15 CU clusters

2nd stage InfiniBand 4x SDR interconnect
(15 sets of 18 links to 8 switches)

IB 4X SDR
Planned Hybrid Compute Node

An 800+ GFlop/s Node!

- Designed to provide one Cell chip per Opteron core

Four Cell Blades

- Four independent IB 4x SDR links
- Passthru modules in Blade Center Chassis

x3755 Opteron node (8 cores)

- Technology & power constrain the design for a 2008 delivery timeframe

IBM

NNSA
Planned Hybrid Compute Node

An 800+ GFlop/s Node!

Designed to provide one Cell chip per Opteron core

Four Cell Blades

Four independent IB 4x SDR links

(1 GB/s) (~5 us)

IB 4x SDR (1 GB/s)

To cluster fabric

x3755 Opteron node (8 cores)

Technology & power constrain the design for a 2008 delivery timeframe
New Improved Roadrunner final system is better

- Keep current base system
  - 70 TF capacity resource in secure
  - Fully available for stockpile stewardship
  - No restabilization after Cells added

- “Next generation” PetaFlop system on same schedule
  - Still based on existing technology
  - Better performance
  - PetaFlop/s demo earlier
  - “Science runs” in open now possible
New hybrid compute node is much improved

400+ GFlop/s Performance

Still one Cell chip per Opteron core

Two Cell Blades

Opteron Node (4 cores)

2nd Generation IB 4x DDR
(2 GB/s)

To cluster fabric

4X BW:Flop
2-3X latency

4X BW:Flop
4-5X latency

Improvements
New Improved Roadrunner Phase 3

“Connected Unit” cluster
192 Opteron nodes
(180 w/ 2 dual-Cell blades
connected w/ 4 PCIe x8 links)

~7,000 dual-core Opterons
• ~50 TeraFlop/s (total)

~13,000 eDP Cell chips
• 1.4 PetaFlop/s (Cell)

CU clusters

2nd stage InfiniBand 4x DDR interconnect
(18 sets of 12 links to 8 switches)

2nd Gen IB 4X DDR
Roadrunner metrics for success

- Provide a large “capacity-mode” computing resource using “base” machine for LANL weapons simulations in FY07
  - Unclassified prep work on system & available codes is done
  - Couple of weeks away from classified operation

- Provide a petascale-class architecture in FY08
  - Performance
    - Future workload
      - Faster simulations
      - Better simulations
    - Expected Linpack ≥ 1.0 PF sustained
  - Usability and manageability
    - System management and integration at scale
    - API for programming hybrid system
  - Technology
    - Delivery of advanced technology

Assessment Review in October 2007 for Phase 3
Programming Roadrunner
Roadrunner Hybrid Environment & APIs

• Remote Communication Library (DaCS)
  – Data Communication & Synchronization
  – Process management & synchronization
  – Topology description
  – Error handling
  – OpenMPI has proven useful for early “remote computation” prototyping

• Parallel Computational Framework Library (ALF)
  – Accelerator Library Framework
  – Data partitioning
  – Task & work queue pipelining
  – Process management
  – Error handling
  – Libspe alternative (low-level SPE functions)
Accelerator Library Framework & Data and Communications & Synchronization Library

- Designed by IBM & LANL to be HW agnostic
  - multicore/GPU/Cell, interconnect, even possibly cluster-wide
  - desire technical community participation to extend range
- ALF for Cell is out in SDK 2.1; DaCS in a future SDK
Programming Approach

- MPI for cluster-wide message passing still used at node level
  - Global Arrays, IPC, UPC, GAS languages, etc. also remain possible choices
- Split off computationally intense operation(s) for Cell acceleration
  - This is equivalent to function offload
  - Amdal’s law applies for speedup! You can’t ignore it.
  - Create many-way parallel work units for SPMD on the SPEs
    - MPMD, RPCs, streaming, etc. are also possible
  - Opteron would typically block, but could do concurrent work
  - Embedded MPI communications are possible via “relay” approach

Considerable flexibility and opportunities exist
MPI programs can evolve

- Key concepts:
  - Pair one Cell core with one Opteron core
  - Move node-level large-grain compute-intensive function/algorithm & associated data onto Cell PPE
    - Similar to client-server, RPC, function offload, etc.
    - Can be implemented one function/package at a time
  - Identify & expose many-way fine-grain parallelism for SPEs
    - Add tiling, work decomposition, etc. (good for multicore as well)
    - Design for overlapped memory transfers (prefetch/write-behind) while computing (work queues)
    - Enable 128-bit wide vector processing & alignments (similar to SSE)
  - Retain main code control, I/O and MPI communications on host CPUs
    - Use “message relay” to/from Cells to cluster-wide MPI on host CPUs
If no MPI communications are needed during Cell computation

- **Host CPU**
  - upload
  - download

- **Cell PPE**
  - work tiles

- **SPE**
  - Overlapped DMAs and compute
  - Get
  - Switch buffers
  - Get (prefetch)
  - Wait
  - Compute
  - Put (write behind)
  - Wait
  - Switch buffers
  - Wait
Add “relay” of DaCS ⇔ MPI messages
Programming a hybrid computer

- Decomposition of an application for Cell-acceleration
  1. Opteron code
     - Runs non-accelerated parts of application
     - Participates in usual cluster parallel computations
     - Controls and communicates with Cell PPC code for the accelerated portions
  2. Cell PPC code
     - Works with Opteron code on accelerated portions of application
     - Allocates Cell common memory
     - Communicates with Opteron code
     - Controls and works with its 8 SPEs
  3. Cell SPE code (8-way parallel)
     - Runs on each SPE (SPMD) (MPMD also possible)
     - Shares Cell common memory with PPC code
     - Manages its Local Store (LS), transferring data blocks in/out as necessary
     - Performs vector computations from its LS data
And what about applications
Radiative Heat Transfer on GPU → 30x faster

- **time to compute heat transfer viewfactors within Truchas casting simulation code**
  - hemi-cube method
    - 3.4 GHz 64-bit Xeon
    - Chaparral (from Sandia Nat. Lab.)
  - New plane projection GPU method
    - NVIDIA Quadro FX 1400 GPU

- **GPU implementation of new algorithm 30x faster**
  - Including data transfer (upload/download)
  - Was done only once at t=0 because of computational cost
  - Can now consider re-computing viewfactors dynamically during fill!

- **parallel execution on cluster**
Ocean Simulation on GPU ⇒ 30 - 50x faster

- shallow water equations
  - standard form of Navier-Stokes equations used for atmospheric and oceanographic processes
  - 14 variables per grid point, 9-point stencil, non-linear
  - 2000 x 1000 grid - 45 time-steps/sec
    - **NVIDIA Quadro FX 4500** 30x-50x faster than 3.4 GHz Xeon EM64T
Initial target LANL applications for Roadrunner

- **Transport**
  - PARTISN (neutron transport via Sn)
    - Sweep3D
    - Sparse solver (PCG)
  - MILAGRO (IMC)

- **Particle methods**
  - Molecular dynamics (SPaSM)
    - Data parallel CM-5 implementation
  - Particle-in-cell (VPIC)

- **Eulerian hydro**
  - Direct Numerical Simulation

- **Linear algebra**
  - LINPACK
  - Preconditioned Conjugate Gradient (PCG)
Initial Cell results are very encouraging

- **Transport**
  - neutron transport via $S_n$ (PARTISN)
    - Sweep3D – 5x speedup on Cell
    - sparse linear solver (PCG)
  - radiation transport via implicit Monte Carlo (MILAGRO)
    - 10x speedup for opacity calculation on Cell

- **Particle methods**
  - Molecular Dynamics (e.g. SPaSM)
    - 7x speedup on Cell
  - Particle-in-cell (plasma)
Effect of new hybrid node should be dramatic

Sweep3D (POR RR)

- Cannot effectively overlap all communication with computation
- Performance gain of Cell-HPC chip may not help

Sweep3D (improved RR)

- Overlapped communication & computation
- Faster with (same) current Cell chip
- Performance gain of Cell-HPC chip can be utilized
Scientific understanding enabled by RR in 2008

- Example: multi-scale validation of ICF NIF implosion

**Continuum simulation of NIF implosion**
- Hydro, EOS, radiation transport, nuclear reactions
- Spans 100 µm ⇒ 0.05 µm with $10^{10}$ zones

**PIC simulation of TN plasma**
- Macro-particles with TN burn
- Kinetic & collective effects
- Spans 10 µm ⇒ $10^{-4}$ µm w/ $10^{11}$ particles

**MD simulation of hot plasma**
- Electrons & ions with Coulomb potential
- Spans 0.03 µm ⇒ i & e’s w/ $10^{10}$ particles
LANL Roadrunner home page

More information is available at:
http://www.lanl.gov/roadrunner/

Roadrunner Architecture
Other Roadrunner talks
Computing Trends
Related Internet links
Silent Data Corruption (Side Note)

A sustained PetaFlop-year is \( \approx 10^{22} \) or \( 2^{73} \) operations!

A PF for 8hrs is still \( \approx 10^{19} \) or \( 2^{63} \)

Reliability of a single compute node only represents the tip of an iceberg.