LA-UR- 11-02469

*Approved for public release;
distribution is unlimited.*

*Title:* High-Performance Computing at Los Alamos

*Author(s):* Jamaludin Mohd-Yusof

*Intended for:* 2011 LANL-IPGP Geosciences Workshop
April 26-28th 2011, Santa Fe, New Mexico

Los Alamos
NATIONAL LABORATORY
EST.1943

# High-Performance Computing at Los Alamos

Jamaludin Mohd-Yusof (CCS-7)

2011 LANL-IPGP Geosciences Workshop
April 26-28th 2011, Santa Fe, New Mexico

## Abstract

A very brief history of significant machines in LANL HPC is presented, concluding with Roadrunner and the Open Science projects which were performed. An overview of the HPC Categories is presented, followed by the different funding channels. The Red, Yellow and Turquoise networks, with their restrictions are shown, followed by a summary of the compute facilities available on Yellow and Turquoise. The presentation concludes with an overview of the R&D work ongoing at LANL, including cluster development and other HPC topics being studied.

# High-Performance Computing at Los Alamos

Jamaludin Mohd-Yusof

CCS-7



## Introduction

- **LANL has a long history in HPC**
  - Wide variety of systems in the past
    - First Cray-1 vector supercomputer
    - First TM CM-5 (first #1 on top500)
- **Most recently, Roadrunner was #1 on Top500 (now #7)**
  - First petaflop/s machine
  - First hybrid supercomputer
  - First #1 Top500 to also be on Green500 (#7 at that time)
  - 12k nodes, with Cell Broadband Engine (CBE) accelerators
  - 1.376 Pflop/s
  - 98 TB RAM
  - Demonstrated the viability of large accelerated clusters
  - Many lessons learned about coding for heterogeneous systems

## Open Science on Roadrunner

- **Open Science (unclassified work) on Roadrunner resulted in significant breakthroughs in materials, astronomy, and laser plasma science.**
  - Laser Plasma Interaction using the VPIC (particle-in-cell) code, Lin Yin, P.I.
  - Parallel Replica Nanowires using the ParRep-AMD code, Arthur Voter, P.I.
  - Magnetic Reconnection using the VPIC (particle-in-cell) code, Bill Daughton, P.I.
  - HIV phylogenetics using the ML code, Tanmoy Bhattcharya P.I.
  - Roadrunner Universe using the MC3 code, Salman Habib, P.I.
  - Ejecta and Spall using the SPaSM code, Tim Germann, P.I.
  - Direct Numerical Simulation of Reacting Turbulence, Daniel Livescu, P.I.









## LANL HPC Assets

- **HPC systems fall into 3 categories**
  - Capacity Systems are crucial for the "day-to-day" workload of the LANL weapons program; typically these systems are intended to handle a large number of relatively small calculations. These calculations typically have a duration of days to weeks.
  - Capability Systems are crucial for weapons science research and large weapons calculations; typically these types of calculations involve very high geometric and numerical resolution, three dimensions, and/or computationally expensive physics algorithms. These calculations typically have a duration of weeks to months.
  - Advanced Architecture Systems are crucial for applying leading edge computing technology to selected weapons science research and large weapons calculations; typically these types of calculations involve very high geometric and numerical resolution, three dimensions, and/or computationally expensive physics algorithms. These calculations typically have a duration of weeks to months.

## Program Funding / Priorities

- **NNSA Advanced Simulation and Computing Program (ASC) funds resources (equipment and staff) to meet the high performance computing needs of the Nuclear Weapons Program (NWP)**
  - ASC authorizes access to their HPC resources based upon requirements derived from the NWP
  - User community composed of personnel from 3 NNSA laboratories
- **Los Alamos Institutional Computing Program (IC) funds HPC resources (equipment and staff) for open and collaborative science at Los Alamos National Laboratory.**
  - IC provides an equal opportunity for access to those resources, to every scientist and engineer at LANL through a competitive, peer-reviewed proposal process
  - User community consists of LANL personnel and collaborators
  - Institutional Computing is entirely unclassified



## Current Production Systems

### Classified HPC Systems

Redtail[1]—IBM, 64.6 teraflop/s
Hurricane[1]—Appro, 51.2 teraflop/s
Typhoon[1]—Appro, 106 teraflop/s
Cielo[2]—Cray XE6, 1370 teraflop/s (5/2011) (#10)
Roadrunner[3]—IBM (Cell), 1376 teraflop/s (#7)

### Unclassified HPC Systems

Cerrillos[3]—IBM, 152 teraflop/s
Conejo[1]—SGI, 52.8 teraflop/s
Yellowrail[1]—IBM, 4.9 teraflop/s
Lobo[1]—Appro, 38.3 teraflop/s
Turing[1]—Appro, 9.4 teraflop/s
Mapache[1]—SGI, 50 teraflop/s

1- Capacity
2- Capability
3- Adv. Architecture
ASC Funded
Institutionally Funded

## Computing Systems

- **LANL computing systems fall into 3 levels of classification**
  - Depending on the networks access requirements
- **Red Network**
  - Also known as the Secure Restricted Network (SRN).
  - Physically disconnected from any other network.
- **Yellow Network**
  - Also known as the Unclassified Protected Network (UPN), it is the default location for unclassified computer systems at LANL.
  - Protected from unauthorized access from the Internet via a firewall.
  - Access to yellow systems from the Internet requires strong authentication.
  - Foreign nationals require 982CA approval for yellow ***and*** remote yellow access.
  - Available only via SSL-portal from a US Government-owned computer.



## Computing Networks (cont'd)

- **Turquoise:**
  - Also known as the Open Collaborative Network (OCN)
  - Designed to enhance collaboration with external partners.
  - Allows ssh access from the Internet through a proxy, wtrw.lanl.gov.
  - Different (and separate) filesystem structure from that of Yellow network.
  - Important differences between Turquoise and Yellow:
    - no export-controlled source or data allowed on Turquoise
    - VPN not required for access
    - no HPSS
  - Five Linux clusters currently on this network:
    - Cerrillos, Lobo, Garnet, Conejo and Mapache.

Los Alamos NATIONAL LABORATORY EST.1943
UNCLASSIFIED
Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA
NNSA

## Yellow Compute Clusters

| Name (Program[1]) | Processor | OS | SUs or CUs | Nodes per SU/CU | CPU cores per Node / Total CPUs | Memory per compute Node / Total | Interconnect | Peak (TFlop/s) | Storage |
|---|---|---|---|---|---|---|---|---|---|
| Cielito (ASC) | AMD "Magny-Cours" | SLES-based CLE and CNL | 1 | 76 nodes | 16 / 1216 | 32GB / 2464GB[6] | 3D Torus | 11.7 | 144 TB Panasas |
| rr-dev RRP3[2] (ASC) | AMD opteron + Cell BE | Linux Fedora 9 | 1 | 12 | 8 (4xAMD + 4xCell) / 96 | 33GB / 396GB | Voltaire InfiniBand | 5.09 | 144 TB Panasas |
| Saguaro (R) | AMD opteron | Linux BProc | 1 | 64 | 2 / 128 | 4GB / 128GB | Ethernet | 0.307 | NFS |
| Turing TLCC[3] (ASC) | AMD opteron | Linux (Chaos) | 1 | 64 | 16 / 1,024 | 32GB / 2TB | Voltaire InfiniBand | 9.4 | 144 TB Panasas |
| Yellowrail (ASC) | AMD opteron | Linux | 1 | 139 | 8 / 1,112 | 16GB / 2.22TB | Voltaire InfiniBand | 4.89 | 144 TB Panasas |

[1] Programs: IC=Institutional Computing, ASC=Advanced Simulation and Computing, R=Recharge
[2] RRP3 = Roadrunner Phase III hybrid architecture with both AMD processors and also Cell blades
[3] TLCC = TriLab Linux Capacity Cluster
[6] Cielito has 4 *viz* nodes with 64GB memory each



## Turquoise Compute Clusters

| Name (Program[1]) | Processor | OS | SUs or CUs | Nodes per SU/CU | CPU cores per Node / Total CPUs | Memory per compute Node / Total | Interconnect | Peak (TFlop/s) | Storage |
|---|---|---|---|---|---|---|---|---|---|
| Cerrillos RRP3[2] (IC) | AMD opteron + Cell BE | Linux Fedora 9 | 2 | 180 | 8 (4xAMD + 4xCell) / 2880 | 33GB / 11.8TB | Voltaire InfiniBand | 152 | 800 TB Panasas |
| Conejo (IC) | Intel Xeon x5550 | Linux (Chaos) | 1 | 620 | 8 / 4960 | 24GB / 14.8TB | Mellanox Infiniband | 52.8 | 800 TB Panasas |
| Garnet RRP3[2] (IC) | AMD opteron + Cell BE | Linux Fedora 9 | 1 | 12 | 8 (4xAMD + 4xCell) / 48 | 33GB / 396GB | Voltaire InfiniBand | 5.09 | 800 TB Panasas |
| Lobo TLCC[3] (ASC) | AMD opteron | Linux (Chaos) | 2 | 136 | 16 / 4,352 | 32GB / 8.7TB | Voltaire InfiniBand | 38.3 | 800 TB Panasas |
| Mapache (ASC) | Intel Xeon x5550 | Linux (Chaos) | 1 | 592 | 8 / 4736 | 24GB / 14.2TB | Mellanox Infiniband | 50.4 | 800 TB Panasas |

[1] Programs: IC=Institutional Computing, ASC=Advanced Simulation and Computing
[2] RRP3 = Roadrunner Phase III hybrid architecture with both AMD processors and also Cell blades
[3] TLCC = TriLab Linux Capacity Cluster

Los Alamos NATIONAL LABORATORY EST. 1943
UNCLASSIFIED
Operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA
NNSA

## R&D Systems: Darwin

- **Moderate-scale GPU-accelerated cluster**
  - Provides platform for exploring programming models, accelerated applications, system software, and data intensive issues for emerging architectures
- **120 nodes total**
  - 116 are 4-socket, 12-core AMD Magny-Cours
    - 49 with dual NVIDIA 2050 Tesla GPUs
    - 49 with single ATI 9350 Firestream GPUs
    - 18 with single NVIDIA Quadro 5000 GPUs
      - (attached to Powerwall)
  - 4 are 8-socket, 8-core Intel Nehalem
- **Total is ~ 120 Tflop/s, 6.7 TB**
  - Cerrillos is #49 on the Top500
  - Darwin would be top ~100
    - assuming similar efficiency as existing GPU clusters



## R&D in Computing Technologies

- **Computational co-design of applications, algorithms, numerical methods, and technology for predictive science**
- **Advanced computational architectures and exascale computing (with multiple hardware vendors)**
- **Data intensive computing, parallel I/O, data analysis at scale, and visualization**
- **Programming models for emerging architectures**
- **Heterogeneous computing models for high-performance scientific applications**
- **Resilience and fault tolerance at the hardware, system software stack, and application levels**
- **Interconnect software and communication libraries**
- **Memory systems for HPC (collaboration with vendors)**

## Conclusion

- **LANL continues to lead in HPC**
  - Early adopter/driver of new paradigms
- **LANL has significant HPC resources available**
  - Both classified (2.5 Pf total) and unclassified (310 Tf)
- **CCS has the expertise to adapt and enable migration to petascale and beyond**
  - Development of Open Science Roadrunner applications
  - Deployment of GPU (and other accelerator) enhanced clusters
  - Co-design of applications and hardware to enable maximum performance
  - Available and interested in collaborating with applications developers