Scalability-Centric System Design for Large Scale Computing

Prof. Yutong Lu
School of computer science, NUDT &
State Key Laboratory of High Performance Computing
ythu@nudt.edu.cn
Overview

National university of defense technology

~2,000 Teachers
~15,000 Students
... Others
Overview

Supercomputers in NUDT, Changsha, China

- YH-1 1st Chinese Supercomputer (1983)
- YH-2 1st Chinese GFlos System (1992)
- YH-X 1st Chinese TFlos System
- TH-1 1st Chinese PFlos System
- TH-1A 2.6PF Top1
- TH-2 33.86PF Top1

NSCC-Tianjin, 2010
NSCC-Changsha, 2012
NSCC-Guangzhou, 2013
Outline

- Trend of HPC Architecture
- Scalable System Software Design
- Applications
Challenges

PSPR
- Performance
- Scalability
- Power consumption
- Reliability
Trend of Architecture

- Tree carriages of Performance
  - Frequency
  - ILP
  - Parallelism

- Performance = Parallelism
  - ...
  - Year 2010: TH-1A, 4.7Pflops, 7168Nodes, 186,368 Cores
  - Year 2013: TH-2, 54.9Pflops, 16000Nodes, 3,120,000 Cores
  - ...

- Exploit parallelism
  - Longitude (100,000nodes)
  - Latitude (multi/many cores, SIMD, ILP)
Trend of Architecture

- Heterogeneous architecture
  - Some of top-level supercomputers
    - Tiahhe-1A
      - NVIDIA M2050 GPU
    - Tianhe-2
      - Intel Xeon Phi
    - Titan
      - NVIDIA K20X GPU
  - Heterogeneous systems on latest Top500 list
    - #53 /Top500, #24 /Top100, #4 /Top10

- Compute Efficiency
  - More computations per joule
  - More computations per transistor
Many core processor

- Intel MIC
  - >60 cores, >200 threads
  - 1.15 GHz
  - > 1 TFlops performance
  - 512b SIMD

- GPU, NVIDIA Kapler
  - 2688 cores
  - 732 MHz
  - 1.31 TFlops
Trend of Architecture

Tianhe supercomputers

TH-1A
- GPU vs
- Data Parallel
- Simple instruction
  - Limited scheduling
- GPU Direct available
  - ~40% ↑ MPI communication on Tianhe-1A
  - 5% ↑ Linpack
- Steep learning curve
- Supporting
  - Cuda
  - Open CL
  - ...
- 2CPU + 2GPU Linpack ~71%
- Whole system Linpack 56.5%

TH-2
- MIC
- Multi threads & SIMD
- Flexible modes
  - Native, Offload, Symmetric, Shared
- SIMD available
  - ~ 4.5 times speedup on Tianhe-2
- Relatively easy to get started
- Intel Supporting
- 2CPU + 3MIC Linpack ~76.5%
- Whole system Linpack 61.6%
Trend of Architecture

**GPU**
- Computational Chemistry and Biology
- Numerical Analytics
- Physics
- Manufacturing: CAD and CAE
- Oil and Gas
- Defense and Intelligence
- Computational Finance
- Media and Entertainment

**MIC**
- Computational Chemistry and Biology
- Electronic Structure
- Physics
- Computational Fluid Dynamic
- Astrophysics
- Environment
- Oil and Gas
- Computational Finance
Trend of Architecture

Memory Hierarchy

- Performance of CPU ↑59%, Perf of MEM ↑26%

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Register</td>
<td>1 circle</td>
</tr>
<tr>
<td>L1 Cache</td>
<td>3 circle</td>
</tr>
<tr>
<td>L2 Cache</td>
<td>10 circle</td>
</tr>
<tr>
<td>L3 Cache</td>
<td>30 circle</td>
</tr>
<tr>
<td>Local Mem</td>
<td>150 circle</td>
</tr>
<tr>
<td>Non-local MEM</td>
<td>&gt;1500 circle</td>
</tr>
</tbody>
</table>

- Exploit Data Locality, reduce communication and memory accessing
Trend of Architecture

- Memory architecture will be benefited from multiple technologies
  - Deeper memory hierarchy
  - Advanced package technology
    - 3D stack, MCM
  - Optical connection btw chips
Trend of Architecture

Power Consumption

- PW for data moving / 48X PW for data computing
  - MLA inside core: 100PJ
  - Read inside CPU: 4800PJ
  - Data moving btw cores: 7500PJ
  - Data moving btw nodes: 9000PJ

- DTF, reduce 20% power consumption, with 5% performance losing

- Power control applications, power aware, minimum data moving
Trend of Architecture

Interconnection network

- NIC
  - High Bandwidth
  - Multiple Lanes

- Router
  - High radix Vs. Low radix

- Topology
  - N-D Torus Vs. Fat Tree
  - N Dimension Tree

- Optical
  - High BW, Low Latency, EMC

- Cost

- Topo-aware software
Trend of Architecture

- Communication
- Reliability
- Power
- Programmability

Heavy the burden of Software
Software issues

- Scalability
  - How to use the exist systems better
  - How to explore the next generation systems

- Resilience
  - Reduce the CR overhead
  - Lightweight resilience method

- Power Control

- Programmability

- HPC vs Big data
  - Data management and filesystem
## Highlights of Tianhe-2

<table>
<thead>
<tr>
<th>Perf</th>
<th>54.9PFlops / 33.86PFlops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>16000</td>
</tr>
<tr>
<td>Mem</td>
<td>1.4PB</td>
</tr>
<tr>
<td>Racks</td>
<td>125+8+13+24=170 (720m²)</td>
</tr>
<tr>
<td>Power</td>
<td>17.8 MW (1.9GFlops/W)</td>
</tr>
<tr>
<td>Cool</td>
<td>Close-coupled chilled water cooling</td>
</tr>
</tbody>
</table>

### Components
- **Phi** #48000
- **IVB** #32000
- **FT-1500** #4096
- **APM**
- **CPM**
- **ION**

### System Architecture
- **TH-2** (125 x Rack)
- **TH-Net**
- **Frame** (8 x board)
- **Compute board**
- **Hybrid Hierarchy shared storage System 12.4PB**
Highlights of Tianhe-2

Software Stack

- HPC Application Service Platform
- Scientific Data Visualization System
- Cloud Computing Platform

Multi-Domain Framework

- MPI
- GA
- OpenMC
- Tools & Library

- C/C++/Fortran
- OpenMP
- Intel Offload

Hybrid Runtime System

- Resource Management System
- H²FS Parallel File System
- Kylin Operating System

Intelligent Monitor & System Management

PAE

PDE

PSE
Programming model

- Trend of programming model
  - Whole system
    - MPI
    - New Data-driven model
  - Intra node
    - Various
      - OpenMP, Cuda/OpenCL, OpenACC
  - Others
    - PGAS (Global Array)

- Portability
- Performance
- Simplicity and Symmetry
- Modularity
- Compatibility
- Completeness
- Distributed memory
Scalable MPI

- **Performance**
  - P2P: Bandwidth/Latency
  - Collective communication
  - Communicator/Group operations
  - MPI-Init

- **Resource consumption**
  - Memory
  - Network connection

- **Measurement**?
Scalable MPI

- Mem consumption for MPI implementation
  
  \( p: \) System Scale (#rank)
  
  \( M \sim O(p^2) \) -- conventional implementation based table

  \( P=10^3, \quad M=4B\times10^6=4MB \)  
  
  \( *P=10^6, \quad M=4B\times10^{12}=4TB \)  
  
  \( P=10^8, \quad M=4B\times10^{16}=40PB \)  
  
  \( *P=10^9, \quad M=4B\times10^{18}=4ZB \)

- Data structures should be redesigned
  
  - Communicator, RMA window, protocol buffer…
Scalable MPI

- TH-Express2 & TH-Express2⁺
  - Network Interface Chip: NIC
    - 10Gbps × 8lane
    - 14Gbps × 8lane (plus)
  - Network Router Chip: NRC
    - 16 ports, more (plus)
  - Optic and electronic hybrid network
  - Topology: Fat tree → N Dimension Tree
  - Design for extension to 100PFlops
Scalable MPI

Message Passing services over TH-Express

- Galaxy Express (GLEX)
  - Basic message passing infrastructure on network interface
  - User level communication technology
  - User and kernel API

- MPICH-GLEX Design Consideration
  - Protocol: different communication mechanisms exhibit different performance and resource usage
  - Application characteristic: communication mode, such as nearest-neighbor communication
  - Scalability: balance between performance and resource usage
Scalable MPI

- Message passing protocols
- Various protocols in low level with TH-Net
  - **Eager Protocol**
    - Exclusive RDMA Channel
      - Performance oriented
    - Shared RDMA Channel
      - Scalability oriented
    - Hybrid channels
      - Combine application model
  - **Rendezvous protocol**
    - Zero-copy data transfer based on RDMA Get
  - Performance benefit from the neighborhood communication in a number of applications
Scalable MPI

- **P2P Performance**
  - **TH-Express2**
    - MPI P2P Bandwidth: 6.3GB/s
    - Latency: ~2us
  - **TH-Express2+**
    - MPI P2P Bandwidth: 12GB/s
    - Latency: ~1us
Scalable MPI

Collective communication

- MPI interface level
  - NonBlock collective
  - Alltoallv/AllGetherV
  - Group-split

- Implementation level
  - Scalable algorithm
  - Topology aware
  - Hardware offload

Collective offload

- Construct topology-aware algorithm tree dynamically
- Message pass automatically based on the trigger of NIC
- Bypass effect of OS noise
Scalable MPI

- Collective Optimization for Scalability
  - Two-level Collective Operations
    - Intra-node: shared-memory
    - Inter-node: network
  - Adaptive tree structure
    - K-nominal
    - K-ary
    - K is a variable value
  - Optimization based on topology
    - Mapping processes to nodes
Scalable MPI

- Non-stop and fault Resilient MPI (NR-MPI)
  - Application continue execution without being relaunched
  - Failure detection and MPI state recovery done by runtime
  - Data-backup by application-level diskless C/R
  - Reconstruct of MPI communicator and channel
Domain Framework

- Hides parallel programming complexity and the hierarchy of parallel computers
- Integrates the efficient implementations of parallel fast algorithms
- Provides efficient data structures and solver libraries
- Supports software engineering for code extensibility
Dynamic Software

- Application Complexity: Multidisciplinary, Multi-physics, Multi-scale, Multi-method
- Legacy applications: Long term for developing, Expensive, Difficult
- Autotuning the performance
- Dynamic resources requirement and providing
- Topo-aware and Latency hiding
- Resource sharing & Hybrid runtime
- Fault tolerant and Resilience
- Rethink & Redesign the software
Scientific Discovery

- Creative Computing Technology
  - Hardware, system software, algorithm, applications

- Creative Data Processing Technology
  - Data management, Analysis, Visualization

- Big Data come from
  - Experiment
  - Observation
  - Sensor network
  - Simulation

- Challenge of computing/throughput
HPC Vs Big Data

☐ Increasing I/O requirements
  ➢ Large scale Pre/Post data sets
  ➢ Visualization and Analysis
  ➢ Big science with Big data
  ➢ Expected data volume per simulation from ~GB to ~PB, typically ~100 TB

☐ I/O Bottleneck
  ➢ Scalability, Efficiency, Performance, Economic and durability

☐ What’s needed for Parallel IO interface
  ➢ More hints could be expressed
  ➢ More patterns could be supported
  ➢ Interface to application IO library
Scalable IO Structure

IO Architecture on Tianhe-2

- Multiple Layers & Hybrid Storages
  - Local Disk
  - PCI-E SSD
  - Disk Array
- 6400 local Disks
  - Bus attached
- 256 IO nodes
  - Burst: above 1TB/s
  - TH-Express and IB QDR port
- 64 Storage Servers
  - Sustained: about 100GB/s
Scalable IO Structure

- **H²FS: Hybrid Hierarchy File System**
  - DPU, A fundamental unit for data processing, tightly couples a compute node with its local storage
  - HVN, Hybrid, Unified and Isolated dynamic namespace maintained by centralized servers
  - Layered and enriched metadata, I/O hints as high level metadata

- **I/O API**
  - POSIX
  - MPI-IO
  - Extended API, layout and policy guide
  - HDF5 over POSIX and extended API
  - Object API (todo)
Scalable IO Structure

- Multi Modes supported in Customized HVN
  - **Forward Mode**
    - local storage bypassed, forward & aggregate requests
  - **Burst Buffer Mode**
    - Local storage attached as independent buffer for draining burst I/O, transparent data movement
  - **Local Cooperation Mode**
    - Local storage unified with individualize layout, DHT for unique-file, partitioned layout for shared-file, with minimum global storage involved
  - **Fusion Mode**
    - Local Cooperation + Global, single unification namespace of H²IO storage, customized data moving policy
Scalable IO Structure

Contributions of components in H²FS

<table>
<thead>
<tr>
<th>Component</th>
<th>Performance</th>
<th>Scalability</th>
<th>Ease-to-Use</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>User-level Client</td>
<td></td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I/O Path Management</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>RDMA Communication</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Request Scheduling</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Storage Management</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Unified Namespace</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Forward Mode</td>
<td></td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Asynchronous Mode</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Scalable IO Structure

- Scalable I/O operation
  - Aggregate burst Bw > 500GB/s, IOR benchmark
  - Aggregate metadata throughput > 100,000 op/s, mdtest

Local cooperation HVN
Scalable IO Structure

- Evaluation on typical HPC application
  - Geoeast, seismic data processing software
  - MEASTRO, MADBench2, S3D

<table>
<thead>
<tr>
<th>Node Number</th>
<th>Non-HVN (GB/s)</th>
<th>HVN (GB/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MEASTRO (Unique file)</td>
<td>7.28</td>
</tr>
<tr>
<td></td>
<td>MADBench2 (Unique file)</td>
<td>30.43</td>
</tr>
<tr>
<td></td>
<td>S3D (shared file)</td>
<td>5.33</td>
</tr>
</tbody>
</table>

HVN is flexible, more work todo
Scalable IO Structure

- **HPC benefits**
  - Scalable burst BW for typical HPC application
  - Isolated HVN makes data intensive application individualize their optimization
  - Reduced requirements for costly shared storage
  - Scalability, Efficiency, Economic and Ease of use

- **Data processing benefits**
  - Maximum locality, DPU provides opportunity to schedule tasks close to data
  - Single namespace make post-processing easy
  - Reduction of data movement, better support for in-situ data analysis and data in-transit analysis
Different Levels of Performance

- **Peak performance**
- **LINPACK performance**
  - Avg. 80%
- **Gordon Bell Prize performance**
  - ~30%
- **Application sustained performance**
  - <5%~10%
- **HPCG Benchmark**
  - ~1%
Scalable Applications

HCFD: High-Order SimulaTor of Aerodynamics

- WCNS- Weighted Compact Nonlinear Scheme
- Explicit Runge-Kutta

Grid Block Groups

Node Mapping

MPI: Grid Block groups
Offload: In a grid block group (CPU+MIC)
OpenMP: In one grid block
SIMD: slice of one block
Scalable Applications

HCFD: High-Order SimulaTor of Aerodynamics

- Balanced partition between CPU/MIC inside each node
  - MIC: CPU 0.6~0.8
- Hierarchical data partition & communication
- Overlap the communication and computation using pipeline
- Memory & cache optimization
- Offload mem reuse
- Exploit SIMD
Scalable Applications

HCFD: High-Order SimulaTor of Aerodynamics

- CPU+MIC
  - 7168 nodes with 3 mics/node, 1.376 million cores
  - Grid 682.4 Billion

Graph showing time vs. number of nodes for different simulations.
Applications

The University of Hong Kong
Applications

NSCC-GZ

Material

Physics & Chemistry

Manufac
ture

Industry

Energy

Society

Mathematics

Tianhe-2

Environment

Medicine

Biology

Healthy
Applications

- High Energy Density Physics
- Weather & Climate
- CFD
- Seismic data processing
- Bio-information
- E-Gov & Service
Applications

- **Climate**
  - Global shallow water model, #8664, ~1.7 million cores, 77%

- **Physics**
  - Gyrokinetic Toroidal Code GTC, #2048, ~160,000 cores

- **Business Opinion Analysis**
  - 600TB structured/non structured data with micMR (Hadoop over MIC), #1024, 100 Million Rec/day
Applications

- Cardiac subcellular level nanoscale calcium-spread mechanical simulation
  - Explore the pathogenesis of heart disease
  - 4096 nodes with mic, 1.27PF

- Virtual drug screening - molecular docking calculations
  - DOCK6.5
  - 303,826 compounds conformation(specs)
  - 1,100 drug target (pdtd)
  - Over 334 million docking calculation
Applications

- Combustion flow in the turbulent
  - Stability and flame propagation mechanism, combustion oscillation mechanism

- Fast simulation of complex electromagnetic environment
  - FDTD
  - MOM
  - PO
Applications

- The Catalytic Mechanism of Human Oxidosqualene Cyclase
  - QM/MM MD simulation (Qchem-Tinker)

- Study the pathogenesis of Flavobacterium
  - Research and product development of the key technology in freshwater fish immune disease prevention and control

- Regional Marine digitizing system
  - Pearl River Estuary South China Sea
Applications

- Neutrino Mass Measurement
  - Simulate 13.7-billion-years cosmic evolution

- High-speed rail tunnel aerodynamic effects

- Shock Wave/Turbulent Boundary Layer Interaction
  - Structural safety of the high-speed aircraft
Applications

- Real-time financial market risk quantification computing

- Sources of air pollution in city
  - Pollutant concentration distribution and temporal trace
Applications

Multi-Scale numerical simulation framework

- Immersed Boundary Method
- Two-fluid Model
- Microscopic kinetic models + LBM.
- Lagrangian-Eulerian-stochastic Method

- Two-fluid Model (SCFT+Reptation)
- Stochastic Entanglement Dynamics
- Lattice Boltzmann Methods (LBM)
- Smooth Particle Hydrodynamics (SPH)

Theoretical approach: SCFT

Kinetics of signalling and metabolic pathways
Coarse-grained Monte Carlo (MC) and Molecular Dynamics (MD) and Non-equilibrium MD
Car-Parrinello MD, Quantum MC
Applications

KylinCloud Cloud Platform

- **Architecture**

- **Features**
  - Customized according to the need of various applications and the arch. of TH-2
  - Provide IaaS and PaaS services to applications with efficient resource management and scheduling mechanisms
  - Provide multiple-level user management and quota management to tenants
  - Provide friendly self-service portal and the statistics, reporting and displaying of the usage of resource
Applications

- E-Gov
- RenderCloud
- micMR
- Video Processing
- Electromagnetic Spectrum Management
Applications

- Need custom hybrid algorithms
  - Performance-oriented programming
  - Communication reduction
  - Architecture aware algorithm
  - Dynamic management of resources at all levels
  - Fault Resilient and Oblivious
  - Rethinking heterogeneous new algorithms at the physics model to maximize the performance

- Application Code
  - Scalability, Maintainable
  - Portable, Productivity
Co-design for Scalable System
Summary

- Use the existing systems better
- Many-core will be the main trend for next generation system
- Interconnection communication is critical
- Hybrid hierarchy IO structure
- System designers and application designers should share the burden of Scalability
- Domain-specific application framework may be helpful
- International collaboration is important
Thanks