Sun-APSTC Initiative in Asia Pacific

Dr Simon See

Director (simon.see@sun.com)
High Performance Computing Technology
Asia Pacific Science and Technology Center
and
Associate Professor
Nanyang Technology University
Key Scientific and Engineering Functions

- Access
- Compute
- Data
- Visualization
**Vertical vs. Horizontal Workloads**

- **Data Size**
  - **Small**: Little, Genomics, Fluid, Structure
  - **Large**: Nano Technology, Finance, Meteorology

- **Compute Intensity**
  - **Little**: Chemistry, Crash, EMD
  - **Huge**: Engine Analysis, Simulation, Automotive EMD Simulation, Noise Analysis

- **Fit for Scalar**
  - 32bit-Cluster

- **Fit for Vector**
  - 64bit Shared Memory

- **Workload Characterization**
  - Courtesy of NEC

- **Technology Areas**
  - **32bit-Cluster**: Little, Nano Technology, Finance
  - **64bit Shared Memory**: Huge, Nano Technology, Finance, Real Time, Local Weather Forecast, Engine Analysis, Simulation, Automotive EMD Simulation, Noise Analysis

- **Applications**
IT Requirements for HPTC

• NO single platform or architecture is best fit for ALL type of applications
The HPTC Architecture Dilemma:
Scale Vertically or Scale Horizontally?

Scale Vertically:
- Parallel applications: OpenMP
- Large Shared Memory
- Top Performance
- Higher acquisition cost
- Lower development and management complexity & cost

Scale Horizontally:
- Serial and parallel applications: MPI
- Throughput
- Lower acquisition cost
- Higher development and management complexity & cost

The Deciding Factor

What do the workloads require?
Cache-coherent shared-memory multi-processors (SMP)

- Tightly-coupled: high bandwidth, low latency
- Large, workloads: ad-hoc transaction processing, data warehousing
- Shared pool processors
- Single large memory

Cluster multi-processor

- Loosely coupled
- Standard H/W & S/W
- Highly parallel (web, some HPTC)

Scale Vertically or Horizontally

Scale Vertically

Single OS Instance

Tightly-coupled: high bandwidth, low latency

Large, workloads: ad-hoc transaction processing, data warehousing

Shared pool processors

Single large memory

Cache-coherent shared-memory multi-processors (SMP)

Scale Horizontally

Cluster Mgmt.

Multiple OS Instances

Loosely coupled

Standard H/W & S/W

Highly parallel (web, some HPTC)

Cluster multi-processor
Vertical vs. Horizontal Workloads

**Scale Vertically**
- Commercial Workloads
  - Large databases
  - Transactional databases
  - Data warehouses
- HPTC Workloads
  - Climate modeling
  - Data mining
  - Signal Processing
  - Cryptanalysis
  - Nuclear simulation
  - Some structural analysis
  - EDA full assembly simulation

**Scale Horizontally**
- Commercial Workloads
  - Web servers, Firewalls
  - Proxy servers, Directories
  - SSL, VPN
  - Media streaming
  - XML processing
- HPTC Workloads
  - Seismic analysis
  - Genomics
  - Computational Fluid Dynamics
  - EDA sub-assembly simulation
  - Some Structural Analysis
  - Crash Testing
Workload Performance Factors

- Processor speed, capacity and throughput
- Memory capacity
- System interconnect latency & bandwidth
- Network and storage I/O
- Operating system scalability
- Visualization performance and quality
- Optimized applications
- Network service availability

#1 issue for real world cluster performance and scaling
Compute Grid Family

- SF15k
- SF6800
- V880
- V60x
- V210
- Blades

Networks:
- Gigabit
- Myrinet
- IB
- Fire Link

Now vs Future
Interconnect Options
Scale Vertically or Scale Horizontally?

Cluster Performance

- **SF15K**
  - Parallel applications: OpenMP
  - Large Shared Memory
  - Top Performance
  - Higher acquisition cost
  - Lower development and management complexity

- **SF12K**

- **SF6800**

- **SF4800**

- **V1280**

- **V880**

- **V480**

- **V480**

- **V210**

- **V60X**

Interdependent Threads

- **Infiniband 800 MB/s**
  - 8 µs latency

- **Myrinet 240 MB/s**
  - 7 - 12 µs latency

- **GBE 100 MB/s**
  - Serial and parallel applications: MPI
  - Throughput
  - Lower acquisition cost
  - Higher development and management complexity

- **Sun Fire Link 4.8 GB/s**
  - < 4 µs latency

The Deciding Factor

What do the workloads require?
HPTC Market: Moving to Clusters

Grand Challenge, Extreme Scale: ASCII, HPCS, Earth Simulator

High Performance Community: Research Capability, Productivity

Technical Computing: Production throughput and embarassingly parallel

Vector systems
SMP systems
PCs, Workstations
SMP Clusters (64 bit)
PC Clusters (32 bit)
“PalmPilot” Clusters!? 

Tera-scale Clusters (64 bit)
Market Trends & Business Climate

- Shift from performance to productivity
- Market shifted to SMPs and now shifting increasingly to clusters
- Proliferation of “industry standard” systems
- Research market often procuring to peak Teraflops (which drives toward commodity)
- Increasing influence of IT over engineering & research end-users (drives toward commodity)
- Linux frequently written into procurements
- Buying cycles stretching due to budget environment and accountability
HPTC: Critical Business Needs

Make the right choice between capability and cost!

- Scalable, heterogeneous, distributed data access
- Scalable end-to-end computing architectures
- Simplified application development and deployment
- Simplified system management and delivery of services
- Scalable and distributed visualization

Provide the required services at an acceptable cost
The Network is the Computer

Evolution in HPTC

1980
1990
2000
2010

- Workstation Computing
- Client / Server Computing
- GRID Computing
GRID as an Operational Concept
Workflow Service Definition & Mapping

Data Storage, Management and Access Services

Capability Computing Services

Capacity Computing Services

Collaboration, Graphics, and Visualization Services
HPTC Strategy

- **Industry Trend: Grid**
  - “All Grid, All the Time”
  - Grid product family
- **Seeds: High-End**
  - HPCS, iHEC
  - Sun Labs interface
- **GTM: Solutions**
  - Intersection of “solutions” and Grid
- **Technology: HPTC Web Services**
  - GGF, OGSA, select ISVs, Web Svcs
Key Scientific and Engineering Functions

- Needs
  - Interactive visualization of extremely large data sets
  - Ability to view data in large-screen collaborative, immersive environment
  - Performance scalability
  - Visual quality
  - Ease of application deployment in multi-display environments
Graphics Grid:
Access for More Users to Visualization Services at Required Visual Quality and Performance Levels

Storage
Compute
Visualization
Display

SAN/NAS
Compute Cluster
Graphics InterConnect
Digital Video Delivery
Visualization Services Over LAN/WAN

Clients
Key Scientific and Engineering Functions

- Needs
  - Improve user access experience
  - Sharing resources
  - Simplify delivery of technical apps and services
  - Collaborative development environment
Sun Collaborative Computing

**Sun Technologies**

- Java™
- Jini™

- Java™
- Jini™

**Constituencies**

- Scientist
- Engineer
- Scientist developer
- Engineer developer
- Service Providers
- ISVs

**Client**

- Sun ONE Studio™ Suite
- Sun ONE™

**Portal**

- Sun ONE Studio™ Suite

**Shared Pool of Resources**

- Virtualization of network, compute, storage resources—N1 initiative
- Resource optimization
- Thin-node and fat SMP clusters
- Single OS (monolithic) scalability
- RAS

- Data Center managers
- IT stakeholders
Grid Evolutionary Strategy: starting with the cluster grid

- Simplest Grid deployment
- Maximum utilization of departmental resources
- Resources allocated based on priorities

Enterprise Grid
- Resources shared within the enterprise
- Policies ensure computing on demand
- Gives multiple groups seamless access to enterprise resources

Global Grid
- Resources shared over the Internet
- Global view of distributed datasets
- Growth path for enterprise Grids
## Grid Infrastructure Software:

### Integrated Stack

<table>
<thead>
<tr>
<th>Web Interface</th>
<th>Global Grid Middleware</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun™ ONE Grid Engine Portal</td>
<td>Avaki</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Systems and Application Administration</th>
<th>Sun ™ONE Studio Development Environment</th>
<th>Distributed Resource Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1™</td>
<td>Compilers and Performance Tools</td>
<td>Sun™ ONE Grid Engine Family</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Operating Environment</th>
<th>Security</th>
<th>Interconnects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solaris™ / Linux</td>
<td>Security</td>
<td>commodity, Myrinet, Sun Fire™Link</td>
</tr>
</tbody>
</table>

- Sun Fire™ SPARC & X86 Servers, Sun Blade™ Desktops
- Choice of Interconnects (commodity, Myrinet, Sun Fire™Link)
One Secure Solaris OE

Government Level Security for 32- and 64-bit Platforms

Assurance Level

Markets

Assurance Level

1980s 1990s 2000s

Trusted Solaris
B1

Military and Intelligence Focus

Solaris OE

Windows

Linux

General Purpose Platforms

EAL4 (B1 & C2)

Military and Intelligence Focus

Commercial and Government

Trusted Solaris Merges with Solaris OE 10

C2

B1
Sun Enabling Technologies

- **Sun™ Grid Engine Software**
  - High-throughput computational capabilities and resource utilization for cluster grids
  - Enterprise Edition adds policy management for shared-ownership campus grids

- **Sun ONE™ Software**
  - Portal Server: Internet services deployment platform
  - Services to quickly, securely deploy technically demanding portals

Supports Solaris, Linux, and Other Operating Systems
Grid Interface:
Access to Web-based HPTC Services

Sun ONE Grid Portal: A Sun PS Offering
Sun’s HPTC Strategy
Asia Pacific Science and Technology Centers
Objectives

- To provide technical and scientific expertise for technical computing business
- To conduct Applied Research in Science and Engineering
Focus Area

- Life Science Computing
- Engineering Computing
- Financial Computing
- Mathematical Science and Modelling

GRID Computing
HPTC Environment
Performance Tuning
Status FY04

- Singapore Center is opened in December 2002
- Japan Center will be announced Soon!
- China Center is still in planning stage
- India and Malaysia Facilities are up and running
Status FY04

- Singapore Center
  - Collaboration and Projects
    - BioBox (in discussion)
      - A/Prof Tan Tin Wee
    - Financial Engineering (still in discussion)
      - Prof Kah and Prof Lee (NUS)
    - Reliable Computing
      - Prof Liew Kim-Meow
    - Game Theory
      - Prof Robert Gay
  - AIST on APGRID
  - MOU with APBioNet
Support to other industries

• Defense
  – Benchmarking

• Manufacturing
  – Porting and Grid Implementation

• Finance and Banking
  – In progress
ApGrid Branch in Sun's Booth
Sun Grid Engine on the ApGrid Testbed

Ultra Enterprise Cluster + Sun Grid Engine (AIST, Japan)

APAN TransPAC 622Mbps x 2

Job submission using Globus

Sun Demo Station Denver, USA
Hardware:
- 2 nodes
- Sun-SPARC Cluster

Operating System:
- Solaris 9

Grid Software:
- DRM: SGE 5.3
- Globus 2.2

Applications:
- Protein Alignment Analysis
- HPC Clustertools 4
- Forte 6.2
- Forte 7
- Forte 8 Beta
- Slps 6
- Gep

Hardware:
- 1 node
- Itanium Cluster

Operating System:
- Linux

Grid Software:
- DRM: SGE 5.3
- Globus 2.0

Applications:
- Protein Alignment Analysis

Hardware:
- 4 nodes
- Sun-SPARC Cluster

Operating System:
- Solaris

Grid Software:
- DRM: SGE 5.3
- Globus 2.0

Applications:
- Protein Alignment Analysis
- Matlab

Hardware:
- 7 nodes
- Sun-SPARC/PIII Cluster

Operating System:
- Solaris/Linux

Grid Software:
- DRM: SGE 5.3
- Globus 2.0

Applications:
- Protein Alignment Analysis
Campus Grid Connectivity
www.ntu-cg.ntu.edu.sg
Design Grid Problem Solving Environment

This project vision provides an Internet portal giving remote access to general analysis and design tools. Our development road map for this takes a route initially for Computational Fluid Dynamic, Finite Element and Multi-disciplinary Analysis and Design Grid Optimisation for a diverse of science and engineering applications.
We present an approach to compute the efficient frontier for portfolio optimization based on evolutionary programming (EP) technique. Our approach relies on multiple EP runs within a search to create the frontier.
USM/UKM Grid

Heterogeneous Inter-Cluster Proposal Setup

- Inter-Campus Connect (Internet)
  - Secure Inter-Campus Connect using Globus Toolkit
- Transfer Queue

UIA Cluster Grid
- Linux nodes
- Solaris nodes

External Cluster Grid
Collaborative R/D

• Grid Computing
  – Development of Tools, scheduling algorithms, middleware
    • Grid IDE, Superscheduler……
  – and large scale implementation
    • NTU-Campus Grid
    • Singapore National Grid
    • Asia Pacific Grid (APGrid)
GriDE Overview

GriDE is an integrated development environment that make it straightforward for scientists and engineers to construct grid applications. It provides friendly tools to access grid resources and makes the development approach easily and fast.

- Portability
- Scalability
- Convenience
- Modular
- Security
- Transparency
GriDE Architecture

NetBeans Tools Platform

- Flow Editor
- Cross Compiler
- Grid debugger
- Performance Turning
- Data Grid Access
- Project Collaboration

GriDE Layer

- Java
- C/C++
- Fortran
- CoGKit
- MPI-G2
- Grid Simulation

Programming languages/ Libraries Layer

- Globus
- SGE
- NimRod

Middleware Layer

- Computing Resources
GUI based Flow Editor

Based on the predefined templates and grid components, the developers can easily define the work flow of their grid applications. It will automatically compiled by the work flow engine which generates the script or source codes to execute or deploy the application to the grid.
Grid Debugger (1)

Grid Debugger includes the tools to browse the grid resources, execute applications and debug on a grid simulation environment.

**Resource Browser**

- easy to explore Grid resources
- Monitor the resource usage
Grid Debugger (2)

Hierarchical Job Submission
– Convenient submit anywhere you can

• Quick Job Submit
• Multiple Job Submit
Grid Debugger (3)

*Job Monitoring*

– Retrieve detail job execution information

– Tracking the execution history
Grid Debugger (4)

**Debugger**

– Execute in the grid simulation environment

– Debug for parallel applications, and grid services
Performance Tuning

- Monitor grid resource usage
- Monitor execution performance
- Dynamically select resources
- Migrate applications between different resources
Cross Compiler

- Compile for multiple programming languages
- Compile for different operation systems
- Compile for different hardware
Collaborative R/D

- Life Science
  - Workflow Engineering
    - Integration of Portal, Life Science Applications and Grid.
    - Life Science Package (BioBox)
  - Algorithms
    - Work with researchers on new algorithms and map them to Sun platform. Optimization and tuning.
What is the BioBox?
• Easy-to-deploy installation package consisting of Sun OS and most popular Biox applications.

• Users who aren't familiar or want to avoid compilation/installation of OS/Biox applications.
Submitting BLAST job using GEP
Application Categories

- Homology & Similarity Search (Blast)
- Sequence Analysis (Hmmer)
- Structural Prediction (Phylip)
- Molecular Imaging/Modeling (NAMD)
- Others (Biojava)
Deployment Architecture of Bio-ClusterGrid

**Bioinformatics Applications**

- Homology and Similarity Search: Biodas, BLAST, FASTA, GlimmerM, Wise
- Sequence Analysis: ACT, ClusterW, EMBOSS, HMMER, Image, T-Coffee
- Structural Prediction: DOWSER, FastDNAml, LOOPP, MapMaker/QTL, PAML, PHYLIP
- Molecular Imaging/Modeling: Artemis, Cn3D, GROMACS, NAMD, NMRView, RasMol, ReadSeq, TribeMCL, VMD
- Others: Biojava, Bioperl, Biopython

**Functional Description**

**Access Tier** - users access via browser (GEP)

**Hardware**
- Sun Ray Thin Client + Sparc Solaris Sunray Server

**Software**
- Sunray Server 2.0
- Or other desktop operating systems

**Resource Management/Portal Tier**
- GEP provides entry point for job submission
- SGE schedules jobs to execution hosts
- SGE monitors job status/completion of jobs
- GEP presents job status/results to users
- SGE/GEP server acts as file server

**Functional Description**

**Execution Tier** (execution hosts)
- Executes jobs sent to it by Grid Master Hosts
- Returns job status/results to Grid Master Hosts

**Hardware**
- SFV210 Sparc Server (2 cpu, 2GB) - multiple nodes

**Software**
- Solaris 9, Sun Grid Engine, Grid Engine Portal
- Customer/ISV applications
- HPC Cluster Tools, SunOne Studio 8 Compiler
Collaborative R/D

• Physical Science and Engineering
  – Workflow Engineering
    • Integration of Portal, Applications, Grid and collaborative environment
  – Algorithms
    • Work with researchers on new algorithms and map them to Sun platform. Optimization and tuning.
Research in Reliable computing

- “Rate of progress toward the **correct answer.**”

- Implications:
  - Measuring floating-point “performance” is problematic.
  - Convergence everywhere to the wrong answer is **not** helpful!
F/P examples /1

Mathematically:

\[ x = (10^{20}) \times (0.1 - 3 \times (0.1/3)) \]
\[ = 0 \]

\[ y = (10^{20} + 1) - 10^{20} \]
F/P examples /2

\[ X = (10^{20}) \times (0.1 - 3 \times (0.1/3)) \]
\[ = -0.7450580746E+12 \]
\[ ? \ 0 \]

\[ Y = (10^{20} + 1) - 10^{20} \]
\[ = 0.000000000000E+00 \]
\[ ? \ 1 \]
Example: Big $\times$ Small

$x = \text{Big} \times \text{Small}$

Big $= 10^{20}$

Small $= 0.1 - 3 \times (0.1/3)$

Right answer: $x = 0$
Ex: Big1 – Big2

\[ y = \text{Big1} - \text{Big2} \]

\[ \text{Big1} = 10^{20} + 1. \]

\[ \text{Big2} = 10^{20} \]

Right answer: \[ y = 1 \]
Using F/P

! Copyright 03/25/2003 Sun Microsystems, Inc.

REAL(4)     BIG,    X,    Y

BIG = 1.0E+20
X = BIG * ( .1 - 3. * (.1/3.))
Y = (BIG + 1.) - BIG
PRINT '('" X = ", E20.10, ", Y = ", E20.10)', X, Y
PRINT '(1X)'

END
Using F/P

! Copyright 03/25/2003 Sun Microsystems, Inc.

REAL(4) BIG, X, Y

BIG = 1.0E+20

X = BIG * ( 0.1 - 3. * (0.1/3.))

Y = (BIG + 1.) - BIG

PRINT '(" X = ", E20.10, ", Y = ", E20.10)', X, Y

PRINT '(1X)'
Interval Definition

• Represented as \([a,b]\) or \([2,3]\)
  
  – A continuous set of numbers bounded by its endpoints \(a \leq b\) (or \(2 \leq 3\))
Interval Definition

• Represented as \([a,b]\) or \([2,3]\)
  
  – A continuous set of numbers bounded by its endpoints \(a \leq b\) (or \(2 \leq 3\))
  
  – Formally:

\[
[a, b] = \bigcup_{a \leq x \leq b} x
\]

\[
= \{ x \mid a \leq x \leq b \} \text{ where } a \text{ and } b \in \mathbb{R} = (-\infty, +\infty)
\]
Basic Arithmetic Operations

\[
[a, b] + [c, d] = [a + c, b + d]
\]

\[
[a, b] - [c, d] = [a - d, b - c]
\]

\[
[a, b] \times [c, d] = \left[ \begin{array}{c}
\min (a \times c, a \times d, b \times c, b \times d) \\
\max (a \times c, a \times d, b \times c, b \times d)
\end{array} \right]
\]

\[
[a, b] \div [c, d] = \left[ \begin{array}{c}
\min (a \div c, a \div d, b \div c, b \div d) \\
\max (a \div c, a \div d, b \div c, b \div d)
\end{array} \right]
\]

given \(d < 0\), or \(0 < a\).
Using F/P + intervals

! Copyright 03/25/2003 Sun Microsystems, Inc.

REAL(4) BIG, X, Y

INTERVAL(4) BIGI, XI, YI

BIG = 1.0E+20

X = BIG * ( .1 - 3. * (.1/3.))

Y = (BIG + 1.) - BIG

PRINT "(" X = ", E20.10, ", Y = ", E20.10)", X, Y

PRINT "(1X)"

BIGI = 1.0E+20

XI = BIGI * ( .1 - 3 * (.1/3))
Results

F/P:

\[
\begin{align*}
X &= -0.7450580746\times10^{12} \\
Y &= 0.000000000000\times10^{00}
\end{align*}
\]

Intervals:

\[
\begin{align*}
XI &= [-0.3\times10^{-14}, 0.3\times10^{-14}] \\
YI &= [-0.9\times10^{13}, 0.9\times10^{13}]
\end{align*}
\]
Interval Benefits

• Good representation of physical reality
  – Visible accuracy and uncertainty information
    • \( x \in [2.3, 3.5] \implies 2.3 \leq x \leq 3.5 \)
  – Measurement error/uncertainty
  – Rigorous error/sensitivity analysis
  – Machine interval: the set of all points therein
    • Intervals are compact sets (or continua)
• Elegant and fast system hardware and software
  – No exceptional events
    • Algebraically closed real and complex number systems
      – No singularities: division by zero
      – No indeterminate forms: \( 0/0 \), \( \infty - \infty \) and \( 0 \times \infty \)
Reliable Computing

– Limited precision in current computing paradigm
  • Imprecise computation
  • Unreliable result
– New Computing Paradigm
  • Interval Arithmetics
– Interval Arithmetics
  • Represent FP in bounded by Interval
– New Algorithm
Other Research Collaboration

- GIS
  - Digital Biologist
- SBS/NTU
  - Bio/Med Grid
- USM/UKM
  - Bio Grid
- BioTec (Thailand)
  - Bio Cluster and BioX Applications
- ASIT (Philippines)
  - Biobox initiative (training)
- University of Hokkiado
- KRIBB (Korea)
- Univ of Queensland (Aus)
Paper Published


Paper Published

• See, S., A Grid-based Technical Computing Portal for MCAE Applications, HPC Asia 2002, Dec 16-19, Bangalore, India


• See, S., Computer, Internet and Mathematics, Proceeding of second East Asia Regional Conference on Mathematics Education and Ninth Southeast Asian Conference on Mathematics Education, 27-31 May, Singapore

• Pok, V.S., See, S., Thng, I., Use of Rate Control and Historical Data to Improve Performance of Servers, European Simulation Multiconference 2002, 3-5. June 2002, FH-Darmstadt, Germany.

Paper Published (cont)


- See, S., Interval Arithmetics for Multidisciplinary Design Optimization, MECHANICS & MATERIALS IN DESIGN (M2D-4), 4th International Conference, Nagoya International Center, Nagoya, Japan, June 5 - 8, 2002

- 5 papers are being reviewed for publication
Collaboration

- Center of Excellence
- Research Collaboration
  - SunLab
  - Engineering
  - Asia Pacific Science and Technology Center
• http://apstc.sun.com.sg
• http://www.sun.com/education
• simon.see@sun.com
• mcwsee@ntu.edu.sg
The Interconnect Effect
Sun Fire Link Scalability versus Gbit Ethernet

POP – 8 node Ocean Models

NAMD – Molecular Dynamics

Sun Fire Link (on the left)
Gbit Ethernet (on the right)