Kasetsart University

Cluster Software Tools: Beauty is in Simplicity

Putchong Uthayopas
Parallel Research Group
Department of Computer Engineering
Kasetsart University, Bangkok Thailand
email: pu@ku.ac.th

Outline

- Introduction and Problems
- Solutions
- Example: SCE project
- Conclusion

Introduction

- Beowulf cluster is now a very popular platform
  - Excellent price/performance
  - Scalable performance: to teraflops on hundreds of nodes
  - large selection of open source software
- What is wrong with this picture?

Problems

- Building and operating large cluster is painful
- Cluster is still difficult to use with very primitive environment
- Why?
  - Large number of tools focus on exploring complex new algorithms rather than simple but work methods
  - Smart design can hurt ☺
    - Can be configured in a zillion unused ways and solving so many problems that hardly occurred
    - Results:
      - Complex to use, maintain, configure
      - Too difficult for users to learn

Beowulf: Power to the people

- Large Traditional Environment
  - Large and complex computing systems.
  - Professional administration, many internal experts
  - Solving large scientific problems
  - Big funding
- New Environment
  - Small to medium size (8-64 nodes) systems
  - Manage by users
  - Solving small, medium problem. Development, education. Industrial use
  - Small funding

- **Fact: only 500 systems in the world have performance more than 68 Gflops (from Top 500 Lists)**
- **Who we should work for?**

Why this is a problem?

- Assumption used to design tools is totally different for small/medium and large system
- Tools design for large system when used with small system will
  - Based on wrong design assumption
  - Having use algorithm that is unnecessarily complex to solve problems that never exist for small system
    - NFS mount

Beauty is in Simplicity

- Simplicity Usage
  - Focus on building only robust set of necessary functionalities
  - Focus on the ease of use and completeness of the implementation
- Benefit
  - Help users goes right to work quickly although system is not absolutely optimal
  - User can improve the system later once they learn more

Beauty is in Simplicity

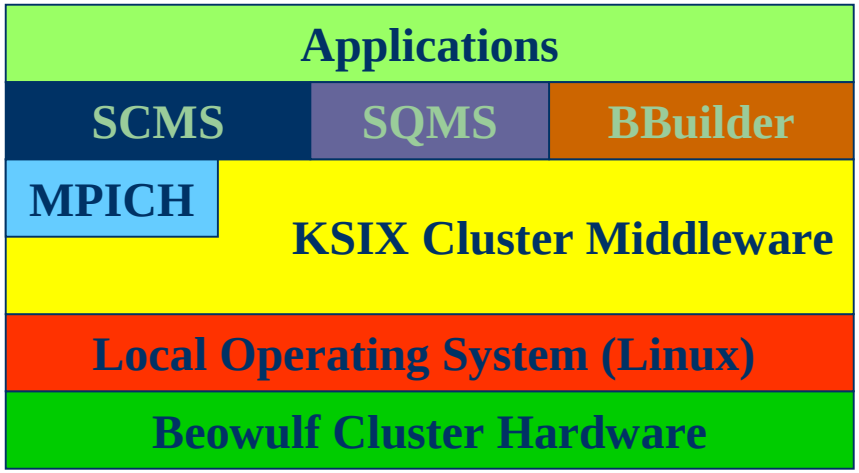
- Simplicity in Design
  - Integration of many simple components
  - Well defined, simple architecture
  - Start simple, extensible by well define API
  - Focus on interoperable and information sharing among components
- Benefit
  - Reduced redundant components. System lighter, faster, and more robust

Our experience

- An integrated cluster software tools
  - Quick starting point for cluster builder
  - Automated cluster builder process up to the point that users can run MPI task with simple batch scheduler
  - Portability : no kernel modification , fully compatible with all software
- SCE components
  - Cluster builder tool (Beowulf Builder/CIMT)
  - Cluster middleware (KSIX)
  - Cluster management tool (SCMS/KCAP)
  - Batch scheduler (SQMS)
  - SCE bundled with fully configured MPICH

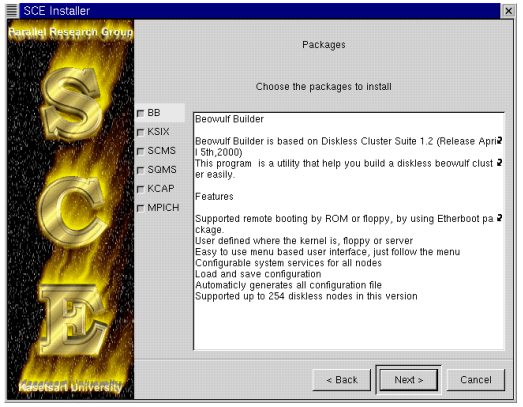
SCE Architectural Concept

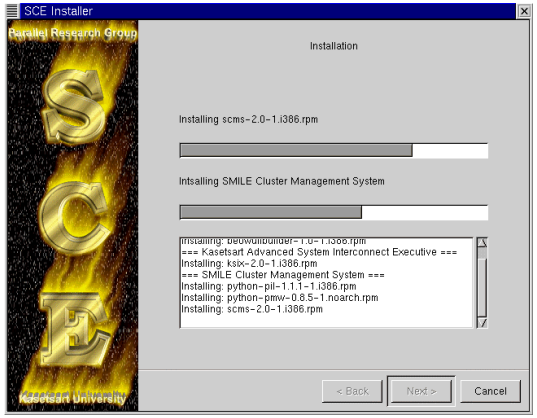
- SCE focus on having simple but extensible architecture
- Offers API and services to ease tool development under SCE
- Single unify configuration service that can be access cluster wide. Shared by all tools



SCE Installer

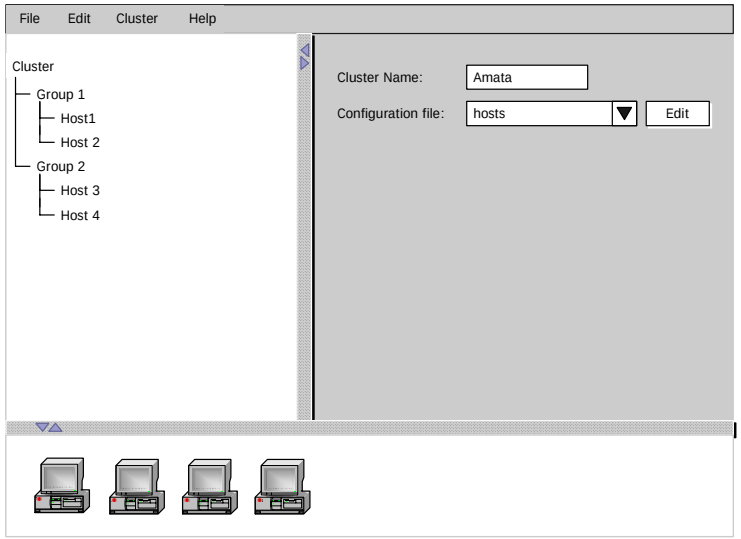
- Automate the installation process up to the point that user can run MPI applications
  - Install package (RPM)
  - Invoke wizard of each package
- Framework that make it easy to integrate more tool into SCE





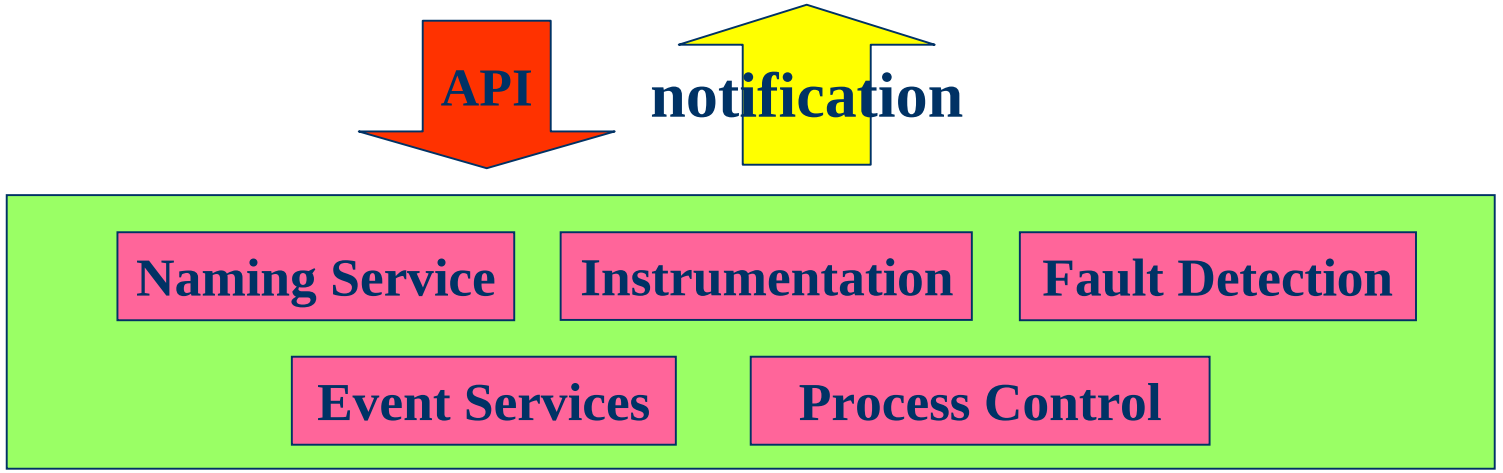
CimT: Cluster Infrastructure Management Tools

- Our next generation tool to build and manage cluster
  - Add,delete,change nodes configuration
- Support both command line and GUI mode
- Support both diskless/diskfull node
  - Support RedHat Linux



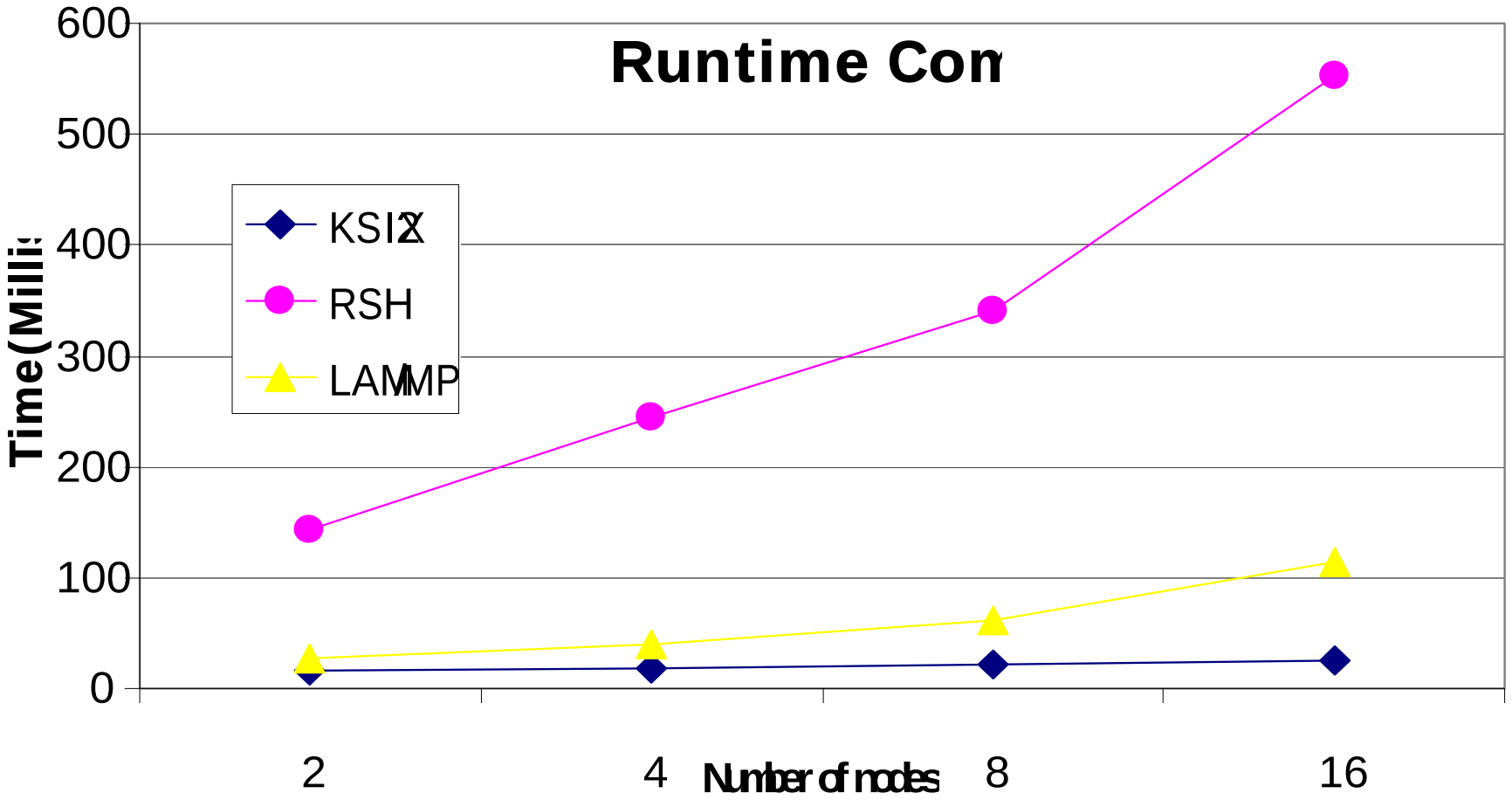
KSIX Cluster Middleware

- Provides many services (about 30 APIs) include
  - Global process management, signal delivery
  - Cluster membership management
  - Distributed event service
  - Naming service
- KSIX must be booted first, follows by the other tools



KSIX Usage

- Speed up parallel unix command with fast process creation
- Process Management for SQMS Batch Scheduler
- Dynamic process creation for MPI (MPICH) (in the future)
- Simple debugging support for MPI

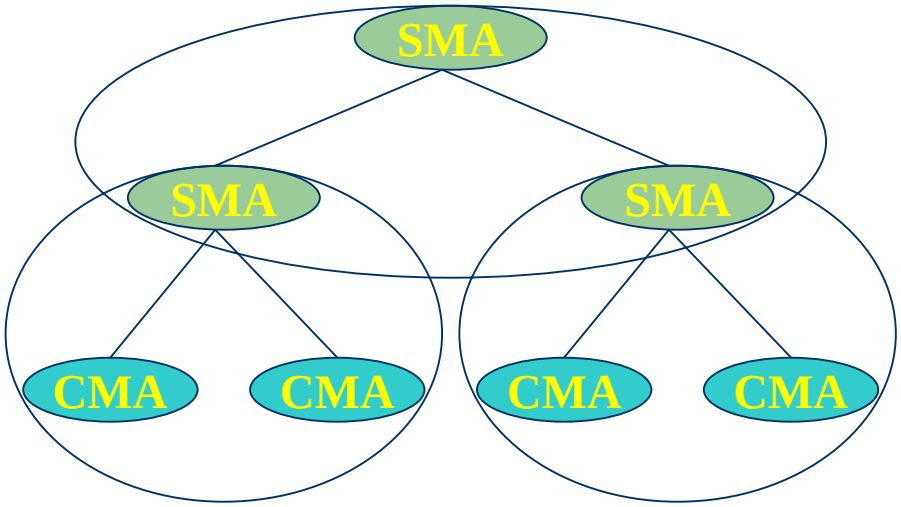

Runtime Con
Time(Milli
600
500
400
300
200
100
0
KSIX
RSH
LAMMP
2
4
8
16
Number of nodes

System instrumentation Services

- Scalable and extensible Instrumentation Infrastructure in SCE
  - Get system information such as CPU,I/O, Memory, network
- System consists of
  - SMA - System Management Agent
    - Response information via RMI
  - CMA - Control and Monitoring Agent
    - Collect information on each node to SMA

Scalable Structure

- Heirachical organization of partitions, highly scalable



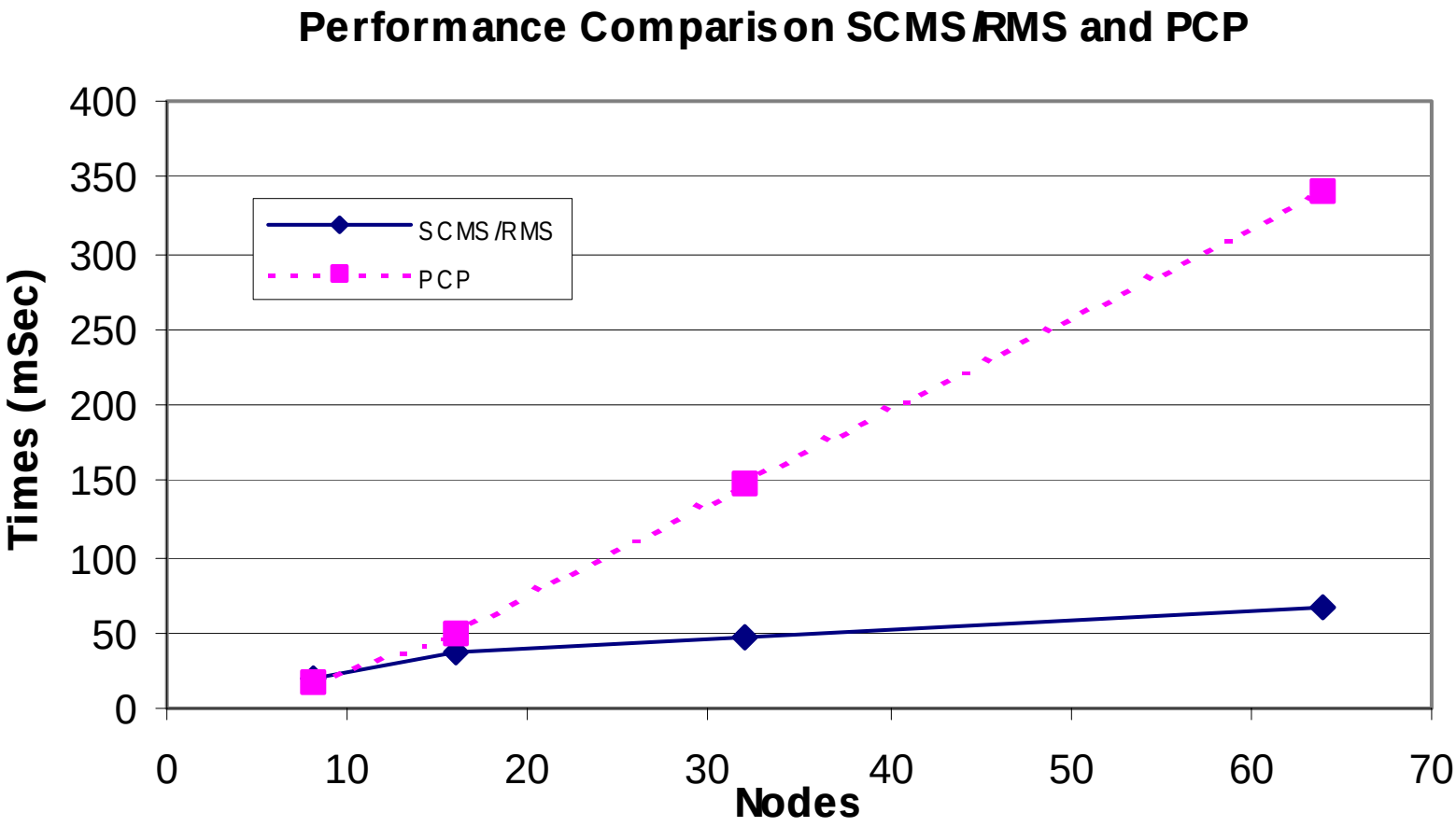
Extensibility

- Separate between /MonitorAnything
  - Information transport mechanism (SCMS/RMS)
  - Interpretation (Client, Agent)
- Loadable Shared Plug-in Module in CMA
  - Define list of plug-in functions for each module
  - Allow initialization and finalization

```
scms_plugin_index _plugin[] = {
    {PLUGIN_CPU,"_hw_cpu_get"},
    {PLUGIN_NULL,""}
}

void _plugin_init(void *args);
void _plugin_exit(void *args);
```

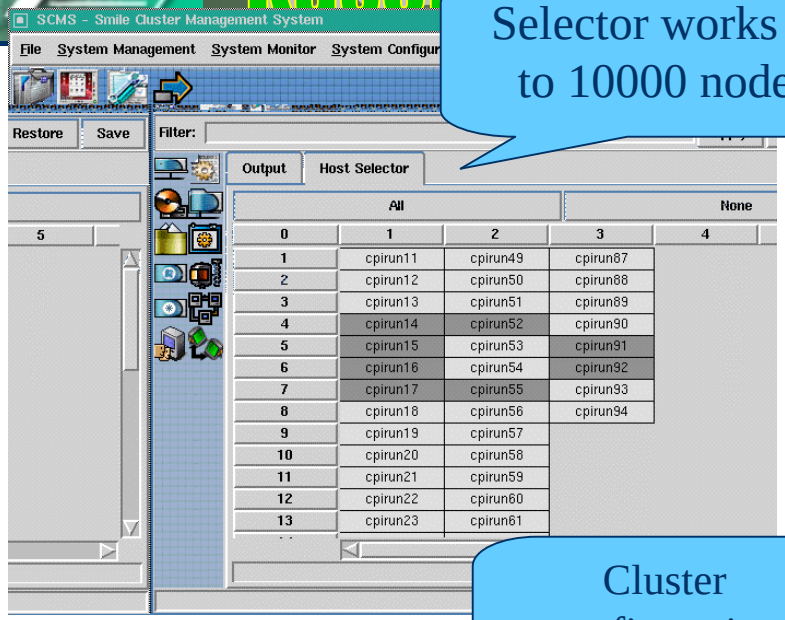
Comparison with Performance Co-Pilot



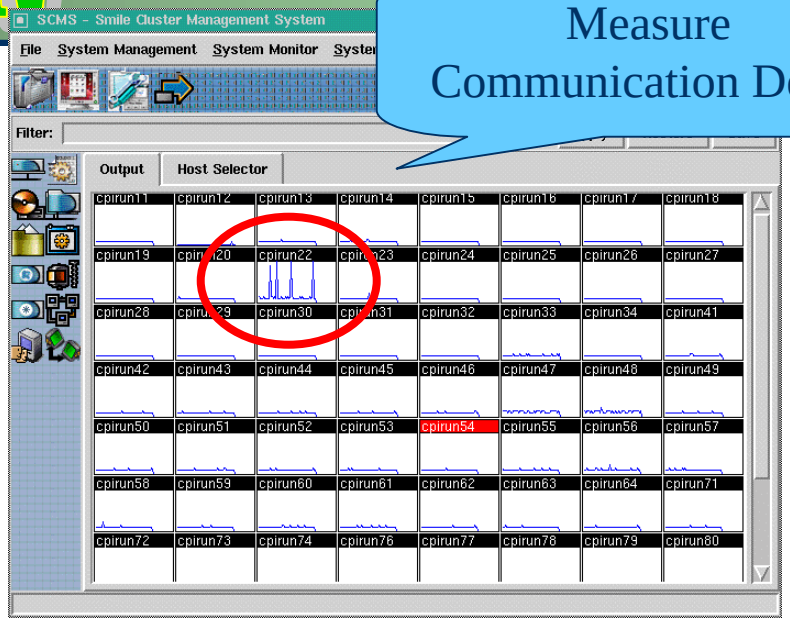
SCMS Cluster Management Tool

- Parallel Unix command
- Complex graphical management console
  - More than 30 operations are supported
  - Configurable look and feel
  - Support large cluster
- Web and VRML Interface for remote monitoring and management

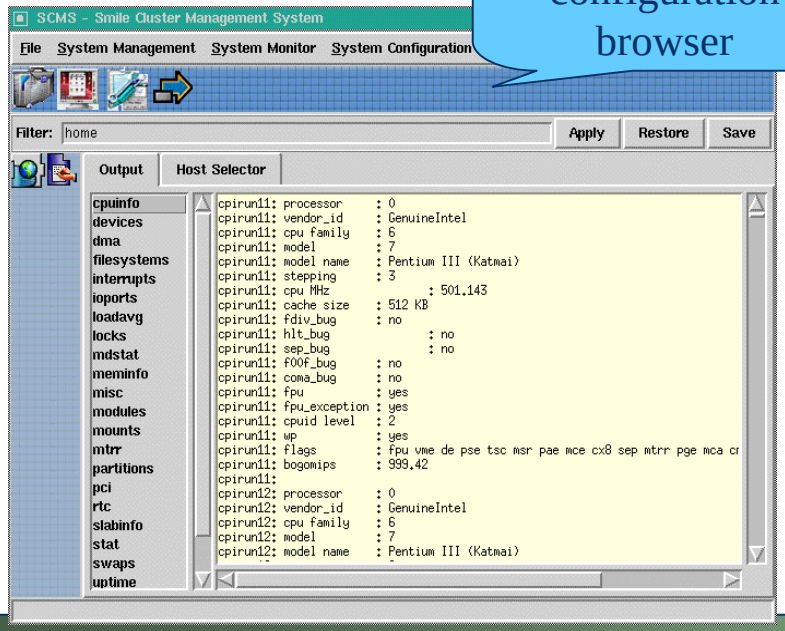
Selector works up to 10000 nodes



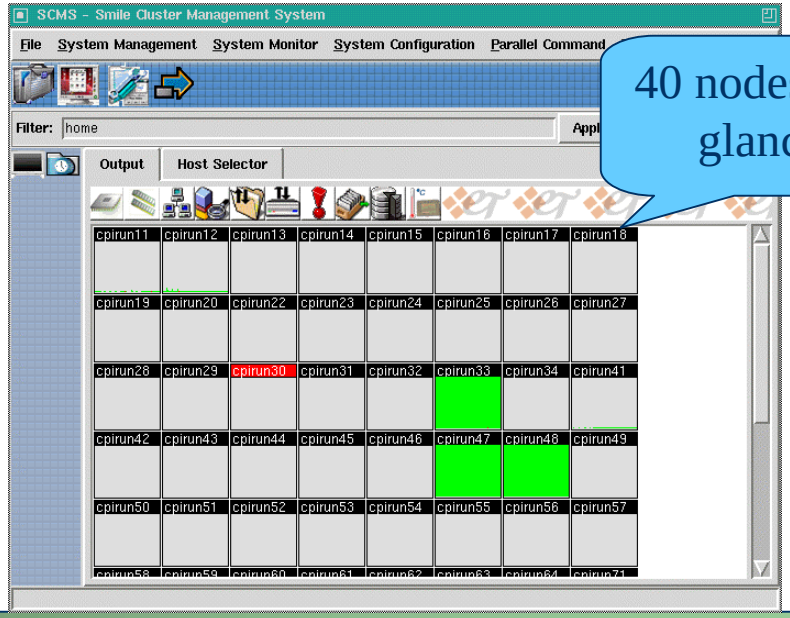
Measure Communication Delay



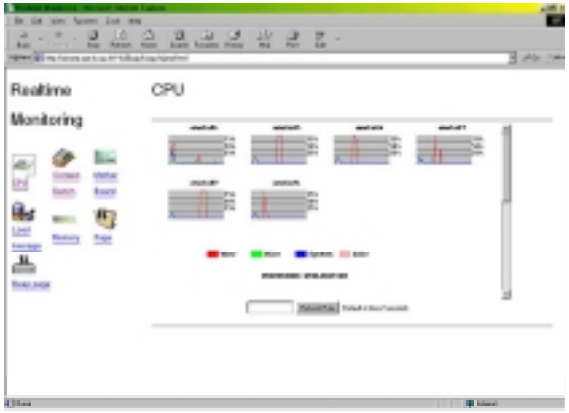
Cluster configuration browser

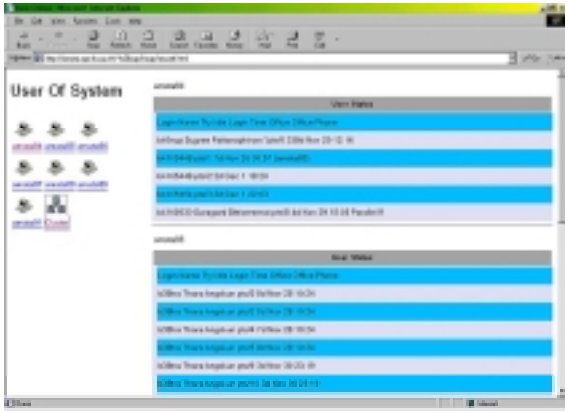


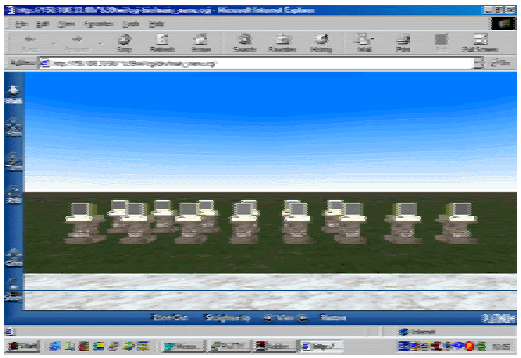
40 nodes at a glance

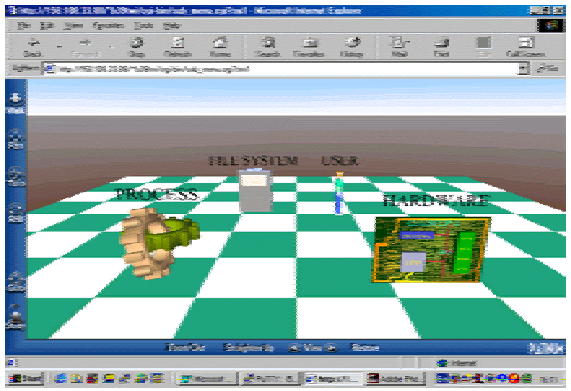


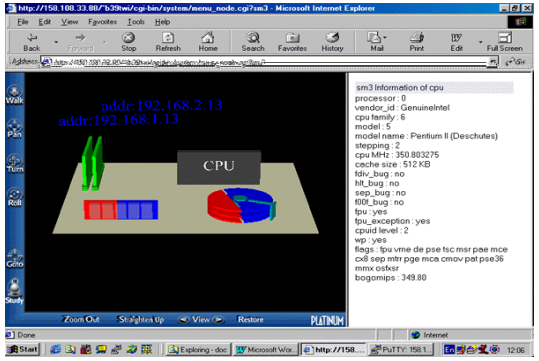

KCAP
ABOUT KCAP

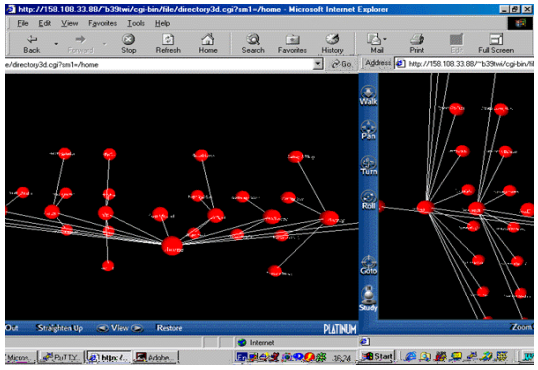

Realtime
Monitoring
CPU

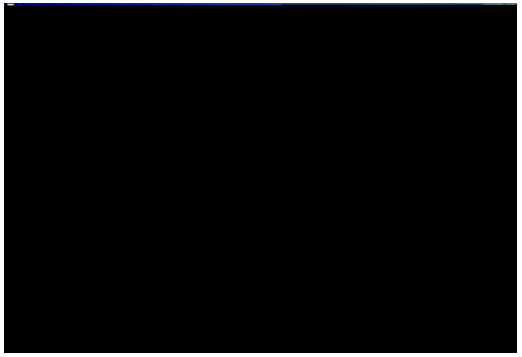

User Of System




FILESYSTEM
USER
PROCESS
HARDWARE

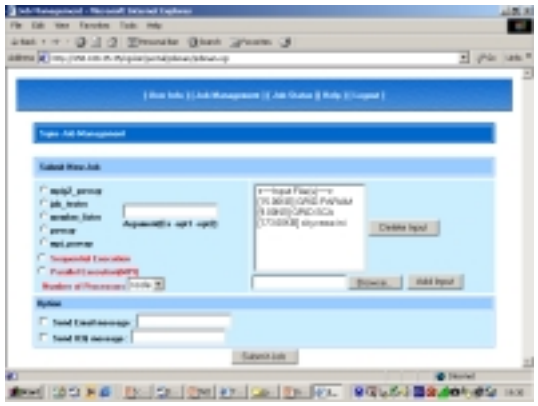

addr:192.168.2.13
addr:192.168.1.13
CPU
sm3 Information of cpu
processor : 0
vendor_id : GenuineIntel
cpu family : 6
model : 5
model name : Pentium II (Deschutes)
stepping : 2
cpu MHz : 350.803275
cache size : 512 KB
fdiv_bug : no
hlt_bug : no
sep_bug : no
f00f_bug : no
fpu : yes
fpu_exception : yes
cpuid level : 2
wp : yes
flags : fpu vme de pse tsc msr pae mce
cx8 sep mtrr pge mca cmov pat pse36
mmx osfxsr
bogomips : 349.80

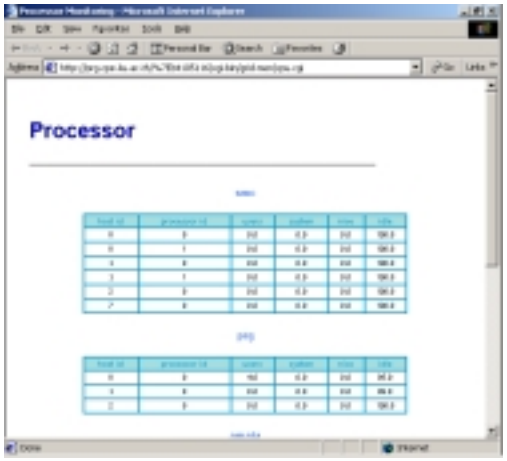


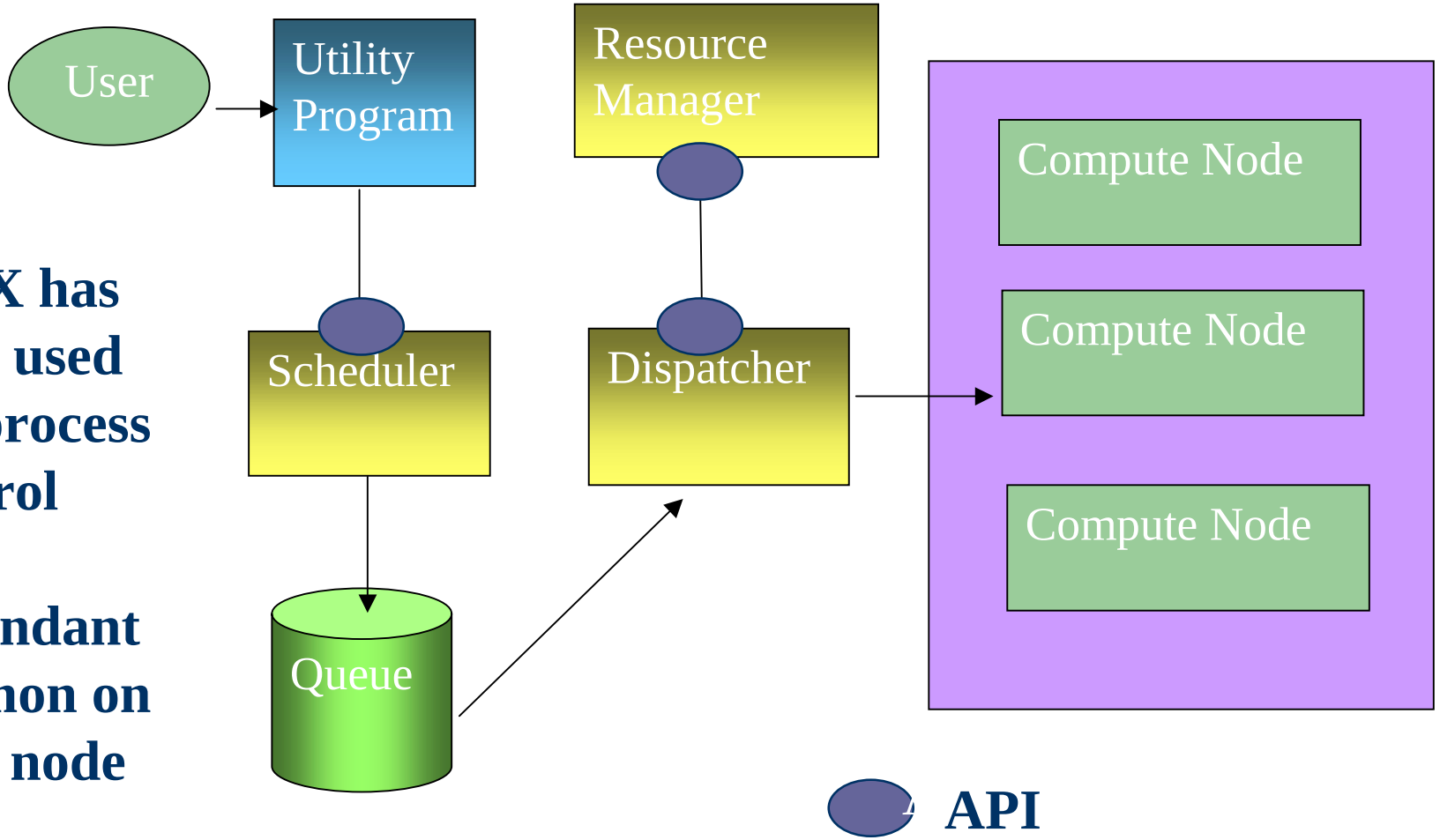


SQMS: Simple Queuing Management System

- Features
  - Sequential task
  - parallel MPI, PVM task
  - Simple load balancing
  - Reconfigurable scheduling policy, resource allocation policy
  - Globus support
  - Web portal support
- Planned for MAUI scheduler support , end of this year



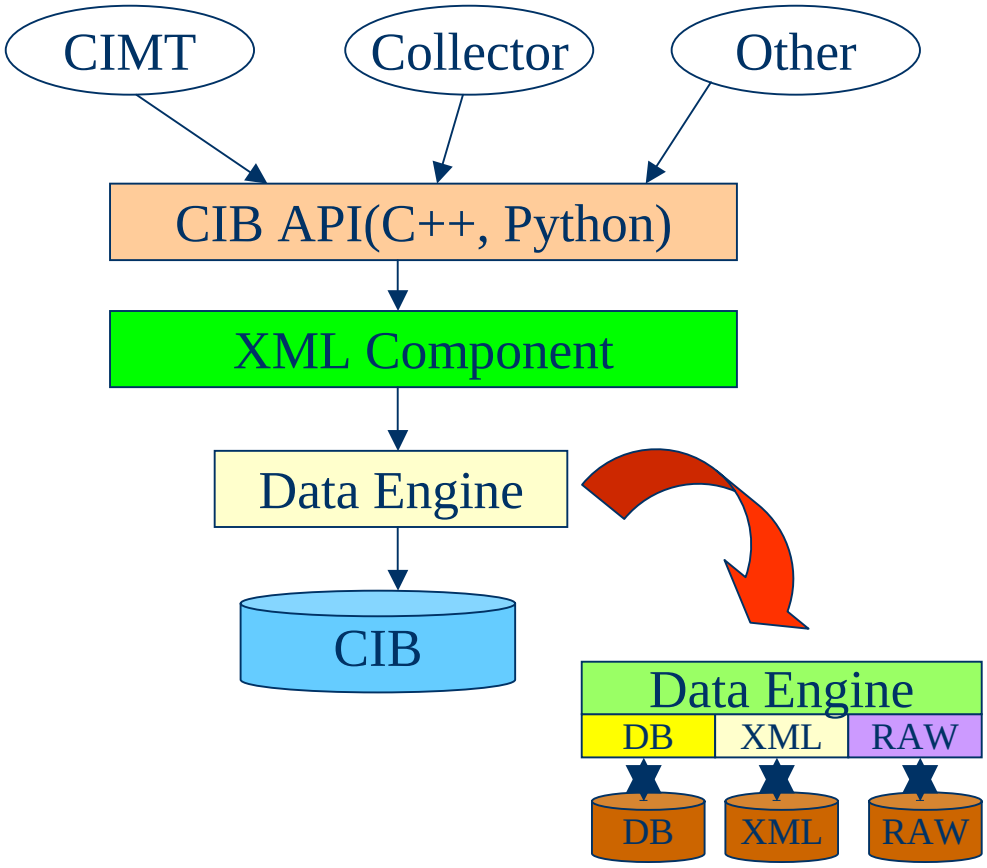




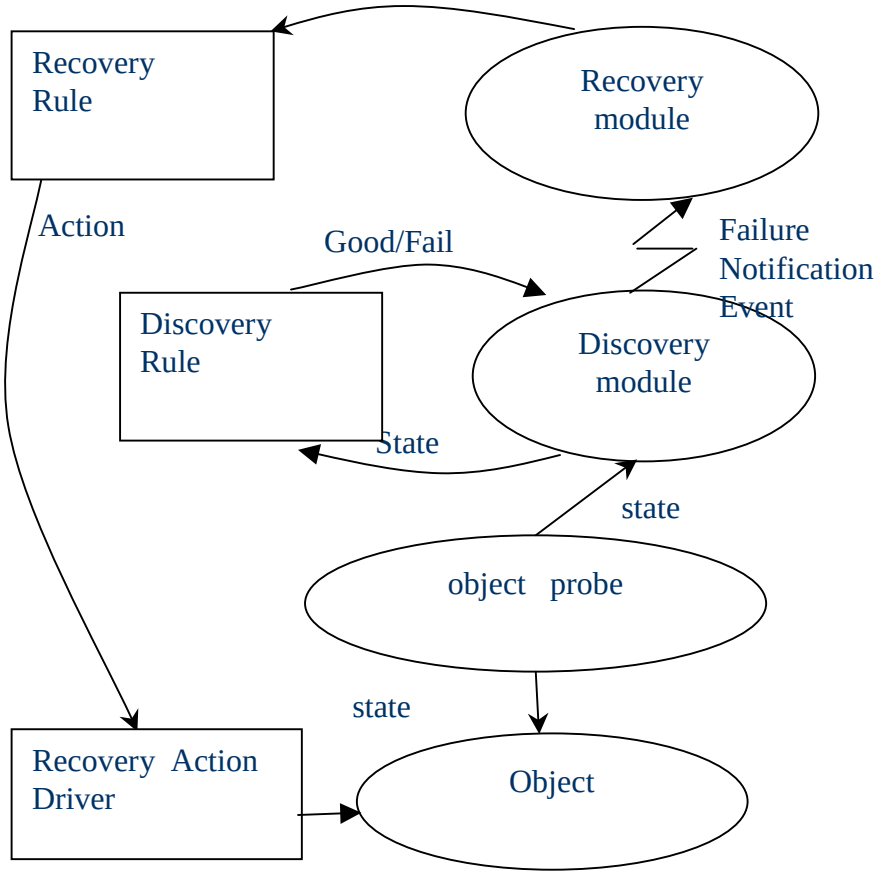
KSIX has been used for process control
No redundant daemon on each node

CIB: Cluster Information Base

- XML based shared data store that can be access by any node store information about
  - current state of cluster operation (dynamic)
    - CPU, Memory used, network traffic
  - System configuration (static)
    - Hardware/ Software configuration



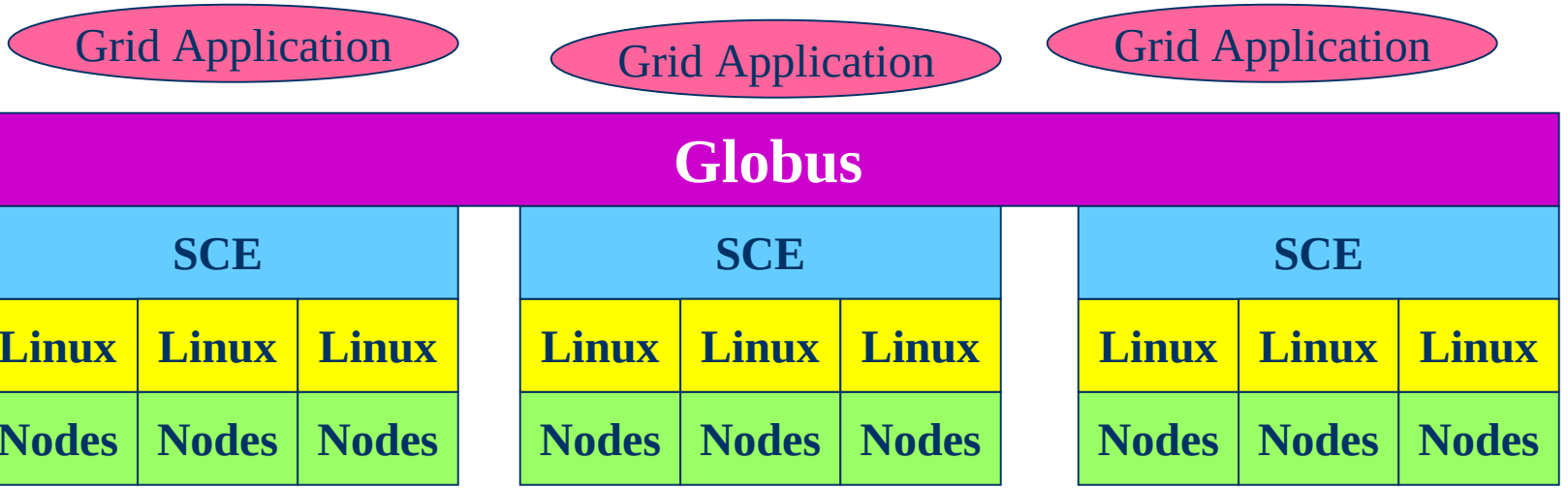
AMATA: Automatic Fault Detection and Recover subsystem



- Subsystem that detect system mal-function
- Send notification
- Allows smart fault discovery and recovery logic to be inserted

SCE and GRID

- SCE has built in Globus support now
  - KSIX is used to start Globus job with our custom GRAM script
  - Grid monitoring is now done using CIB
  - SCE is now used in our THAIGRID project http://prg.cpe.ku.ac.th/thaigrid/



SCE Project Partnership

- AMD
  - SCE is developed on AMD sponsored Athlon/myrinet cluster (AMATA)
  - Partner with us to make and distribute SCE CD to cluster community

- COMPAQ
  - Contribute small alpha cluster for 64 bit port to alpha

COMPAQ

- Terra Soft Solutions
  - PowerPC port on terasoft platform
- SUT, KMITNB, KU (Thailand)

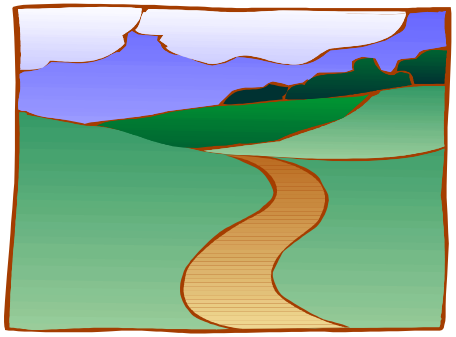
Current Status

- SCE main site is now at

WWW.OPENSCE.ORG

- SCE 1.0 support RedHat 7.1 and 6.2
- SCE 1.2 (released October)
  - Bugs fixed and many new features
- Download
  - http://www.opensce.org
  - http://sourceforge.net/projects/sce

SCE Road Map



- SCE 1.5 (November 2001)
  - Simple CIB
  - More tools and features (such as PAPI/Rabbit support)
- SCE 2.0 (Q1 2002) major change
  - Cluster Information Base (CIB)
  - XML support
  - Working better with MPICH, Globus Grid
  - Smarter Batch Scheduler (SQMS 2.0)
  - Simple debugging and runtime visualization support for MPI applications

Conclusion

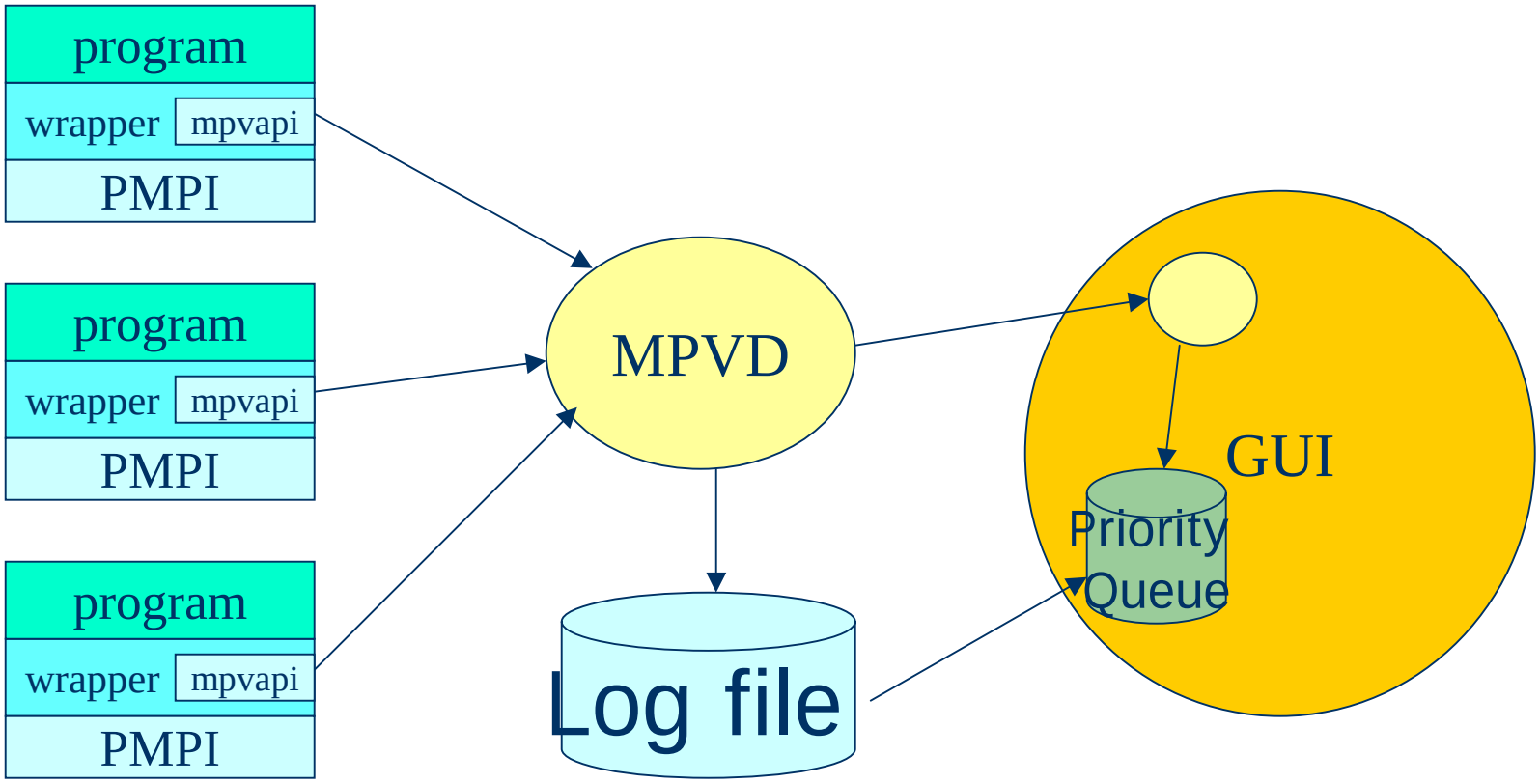
- Proposed design principle for cluster software tools and environment
  - Simplicity
  - Portability
  - Interoperability
  - Extensibility

SPIE

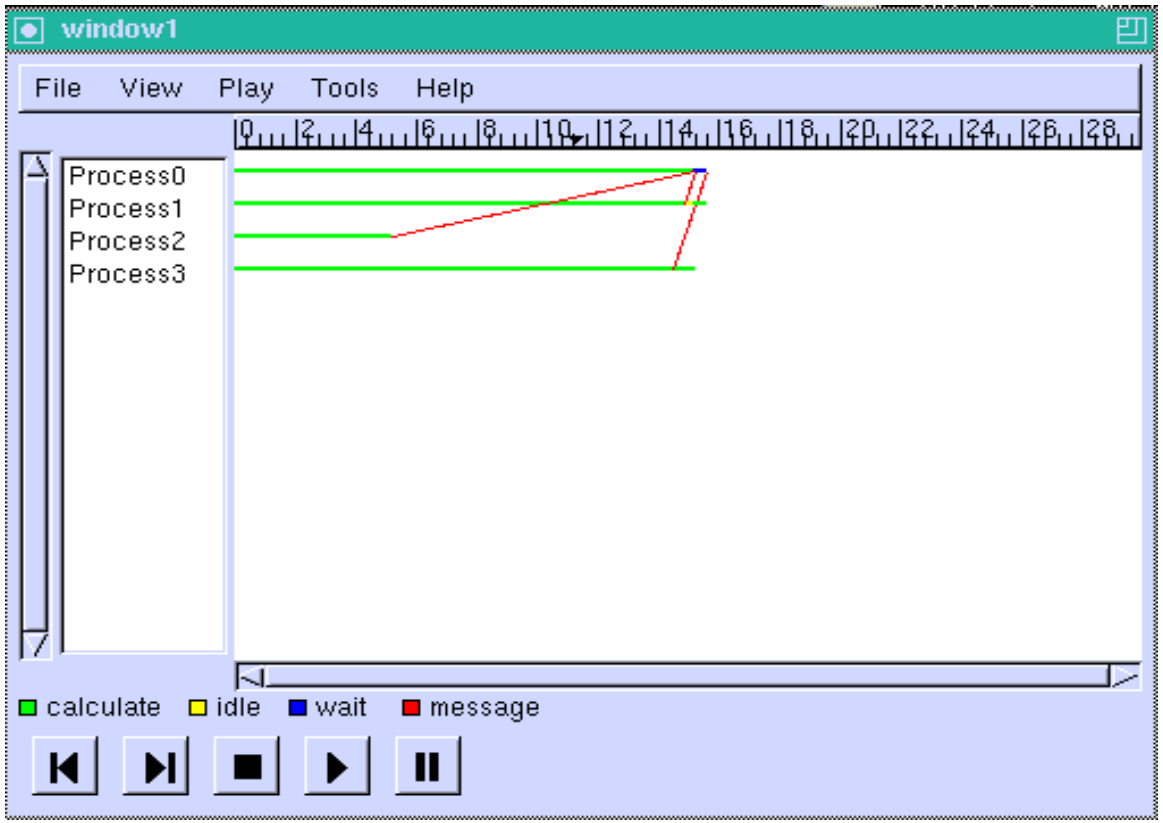
- Software tools still have to evolve to
  - Directly address more realistic requirement
  - More work on user environment


Kasetsart Univers

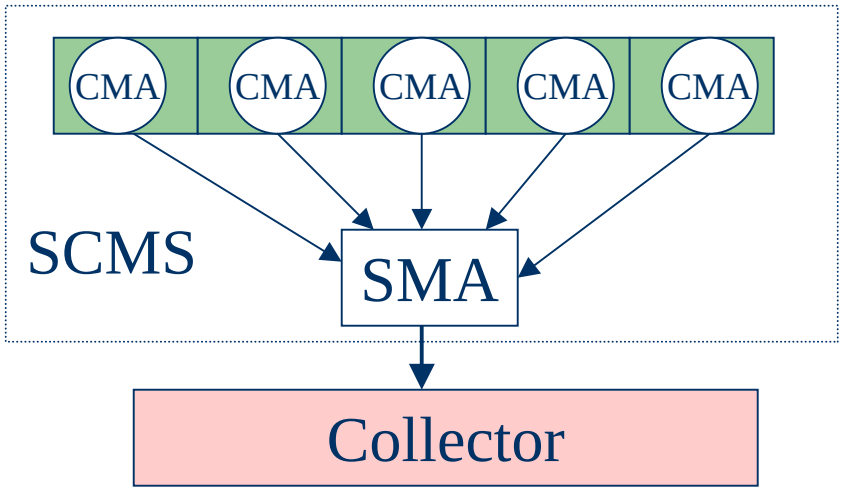
MPVIEW Structure



Result


window1
File
View
Play
Tools
Help
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
Process0
Process1
Process2
Process3
calculate
idle
wait
message

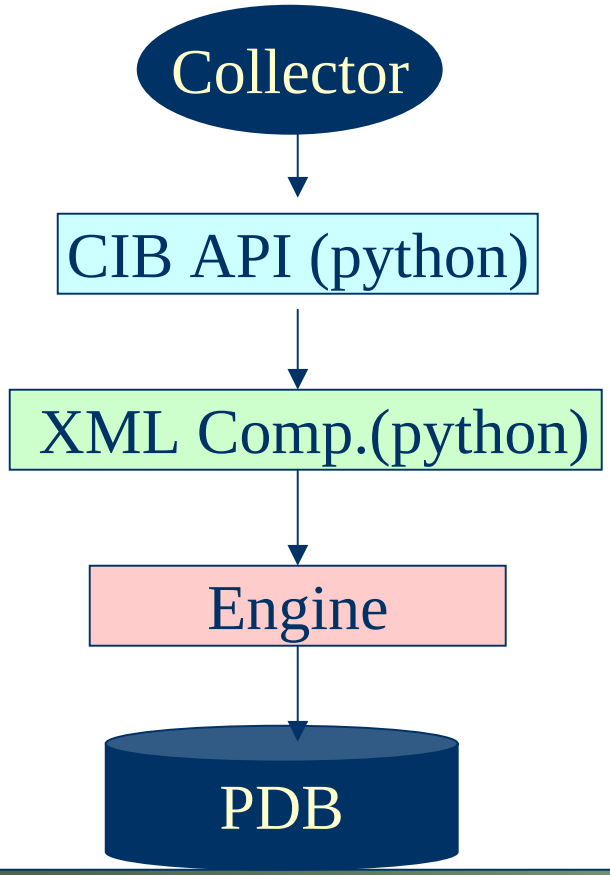
Performance Collector using CIB



- **Collector result example**

{'psi': {'memory': [{'free': 5951488, 'plugin_id': 2, 'shared': 58597376, 'cached': 62603264, 'hid': 0, 'used': 124915712, 'buffers': 17977344}]}, 'psi2': {'memory': [{'free': 25395200, 'plugin_id': 2, 'shared': 66617344, 'cached': 76500992, 'hid': 1, 'used': 105467904, 'buffers': 5656576}]}, 'psi3': {'memory': [{'free': 4124672, 'plugin_id': 2, 'shared': 27926528, 'cached': 95178752, 'hid': 2, 'used': 126611456, 'buffers': 4153344}]}}

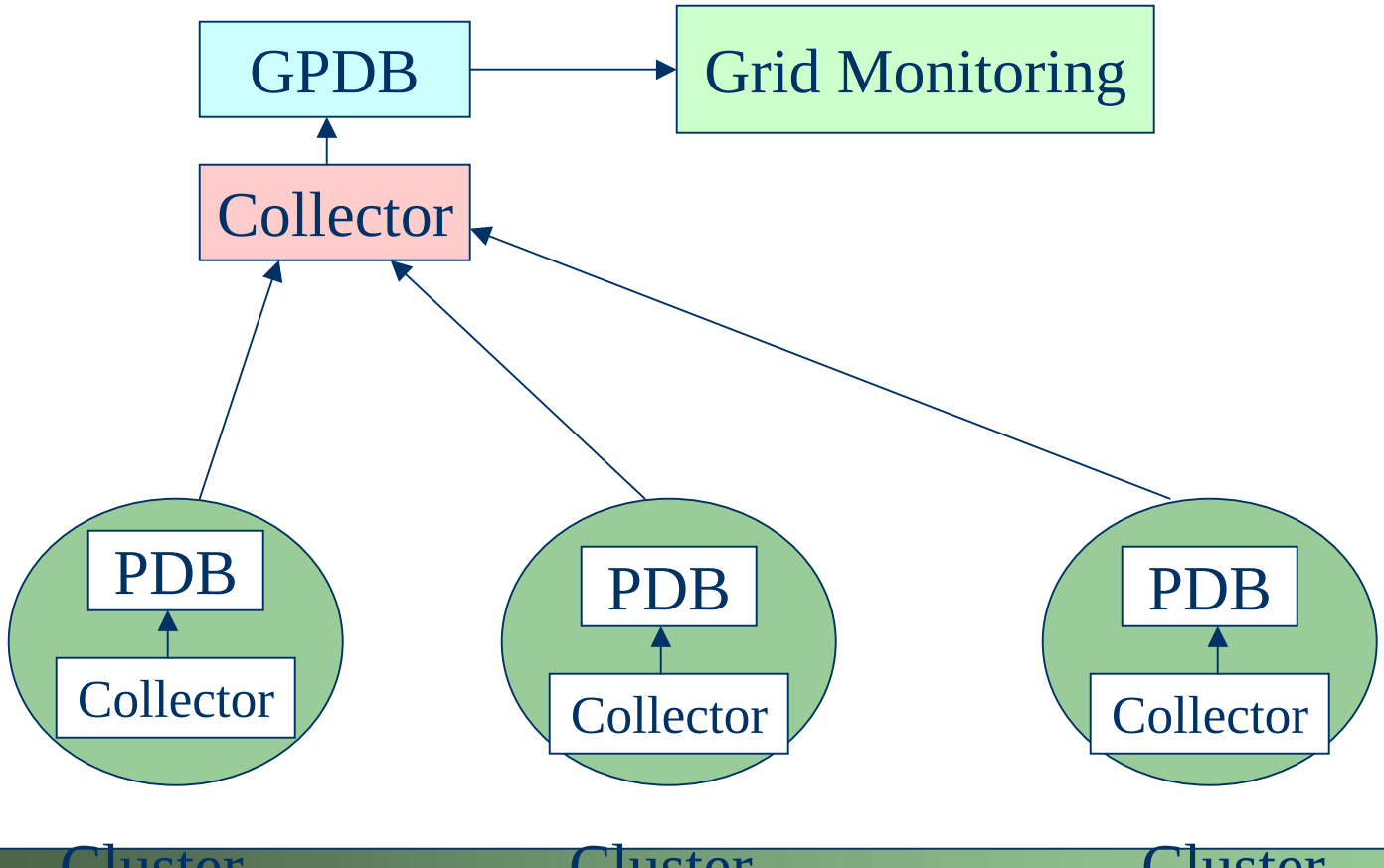
XML Component & CIB



- **Result example**

```
<?xml version="1.0"?>
<cluster>
  <psi>
    <memory>
      <free>5832704</free>
      <plugin_id>2</plugin_id>
      <shared>53731328</shared>
      <cached>61534208</cached>
      <hid>0</hid>
      <used>125034496</used>
      <buffers>24363008</buffers>
    </memory>
  </psi>
</cluster>
```

GRID Extension



SCE Installation