Environment Aware Performance Diagnosis

Karen L. Karavanic
Portland State University
UCSD Performance Modeling and Characterization Lab

Portland State UNIVERSITY

Performance Tuning at the High End

1. Start with coarse-grain view of whole program performance
2. When you see a problem, collect more information to refine this problem.
3. Repeat step #2 until you have a precise enough cause.
4. Collect information to try to refine to particular hosts, processes, modules, functions, files, etc.
5. Repeat step #4 until you have a precise enough location.

This type of iteration can take a user many runs of a program to reach a useful conclusion
--> automated performance diagnosis

Environment-based Performance Problems

Common themes:

- Problem exists because performance differs from an expectation
- Significant time before "key insight" point, with several iterations of optimization and assessment
- Good solutions required "root cause" identification
- Tools gave diagnoses that were too broad, did not identify root causes

Types of Problems:

- Detecting Runtime Interference*
  - ASCI-Q at LANL, IBM SP study at LLNL
- Bug in the Operating System or File System
- Performance Bottlenecks Caused by Faulty Hardware

**Petrini, F., Kerbyson, D. J., and Pakin, S. SC 2003. The case of the missing supercomputer performance.*

**Jones, T. R., Brenner, L. B. and Fier, J. M. Impacts of Operating Systems on the Scalability of Parallel Applications, Technical Report, Lawrence Livermore National Laboratory, 2003.*

Performance Tools Mismatch

Parallel Performance Tools

Paradyn, *TAU*, *Vampir*, *KOJAK*

- Bottleneck detection
- Hardware counters, User-defined metrics

System-Monitoring Tools

NWPerf, *Ganglia*, *OVIS* – Cluster monitoring and analysis

KernInst – Kernel instrumentation

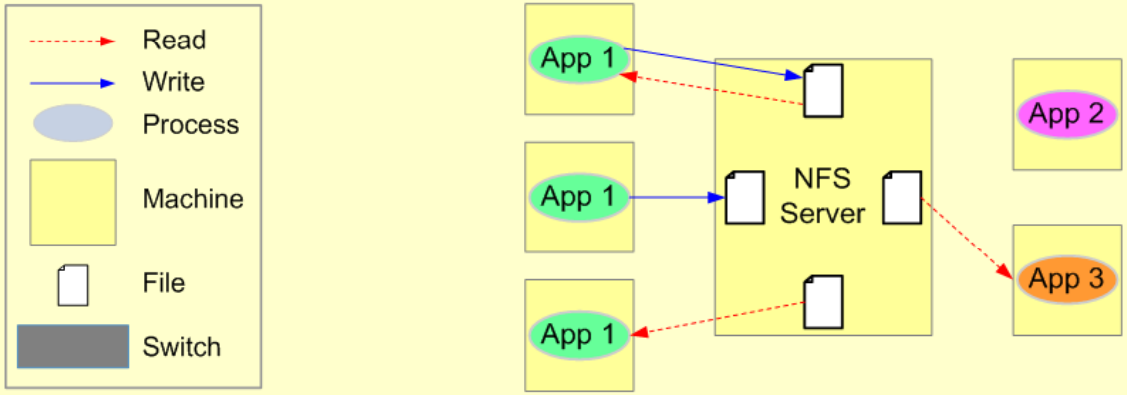
Proc file system, *top*, *vmstat*, *netstat*, *strace* – Single-system

System and Application Tools

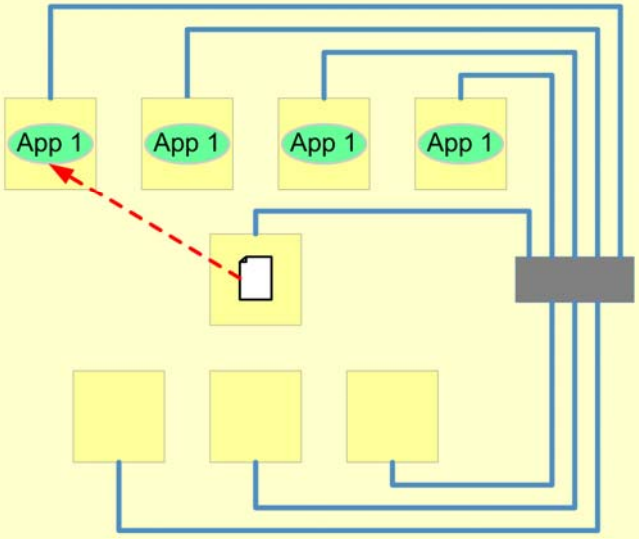
AIX Trace, *Paraver-SCPUs*, *OProfile*, *DCPI*, *CrossWalk*

- Narrow scope
- Post-mortem analysis

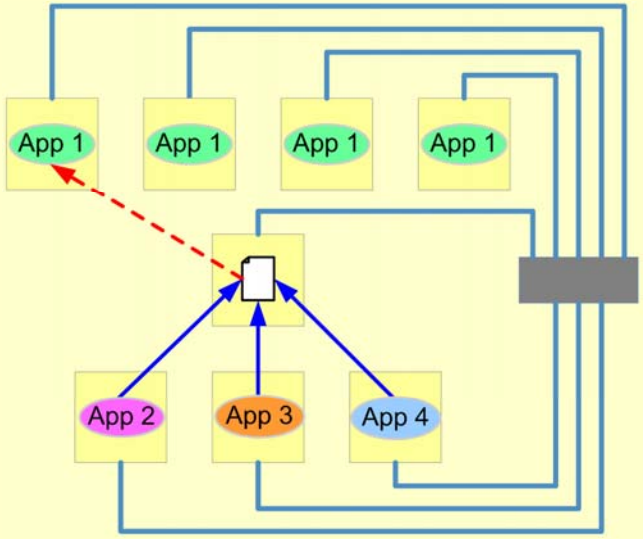
Description of NFS Server Test



General NFS Scenario



NFS Server Test
Normal Environment



NFS Server Test
Suboptimal Environment

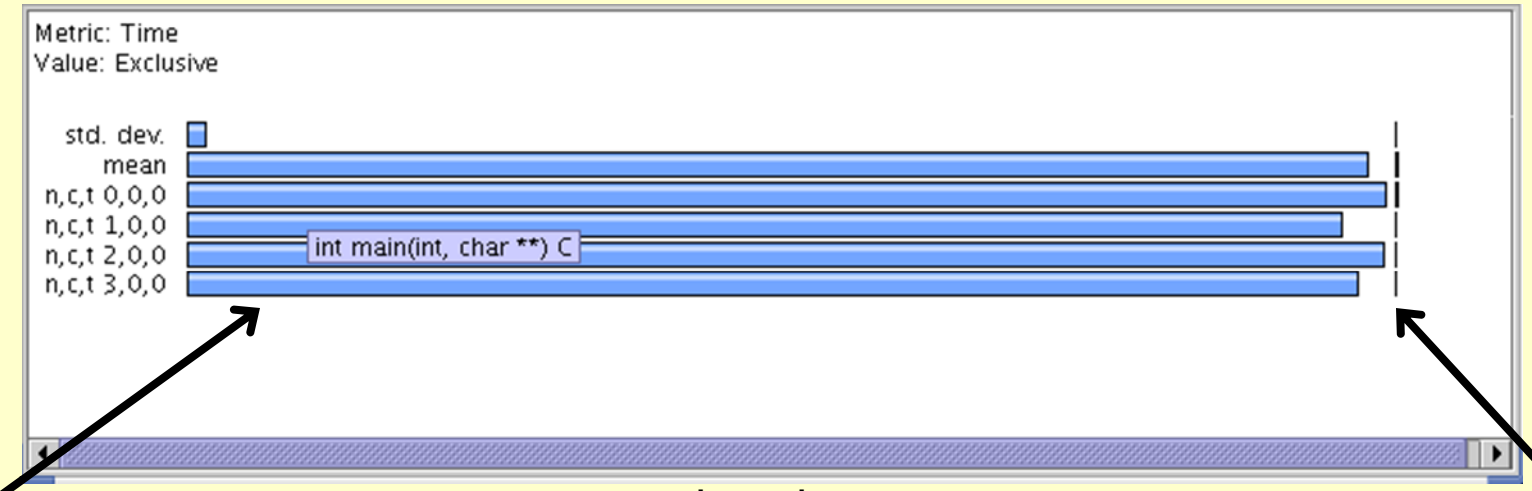
NFS Server Test Results – Application Timing

Timing of Master Process

- Total execution wall-clock time (seconds)
- Wall-clock time for the read operation (seconds)

	Wall-Clock		Read Time	
	Normal	**Suboptimal**	**Normal**	**Suboptimal**
Min.	11.30	11.27	0.00038	0.00036
Max.	12.05	41.28	0.00096	30.00
Avg.	11.48	14.71	0.00051	3.43
Std. Dev.	0.17	8.55	0.00016	8.55

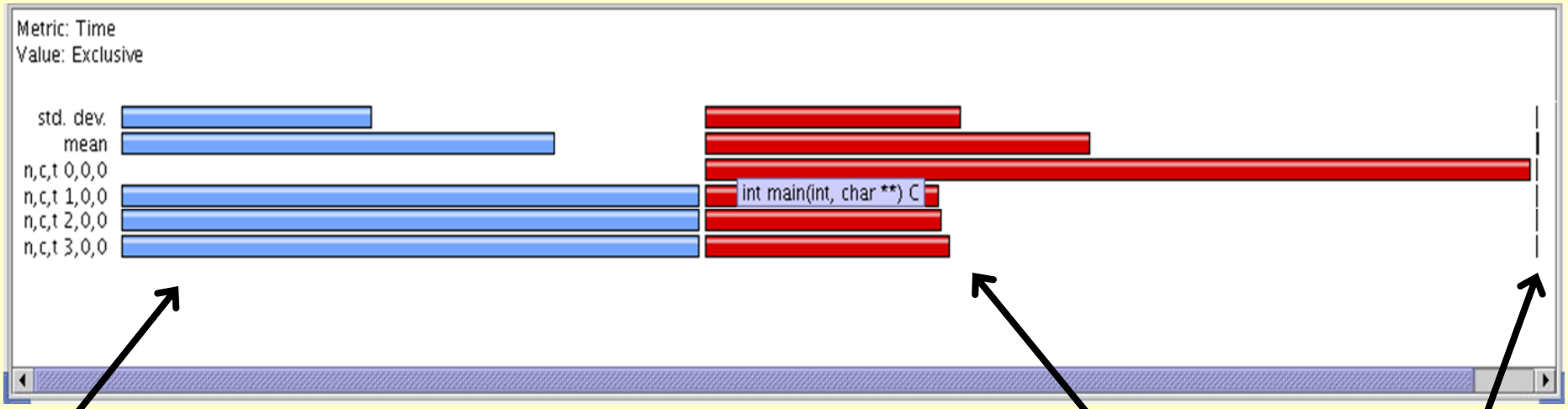
NFS Server Test Results – Application Profile



`main()`

Normal environment

other functions



`MPI_Bcast()`

Suboptimal environment

`main()`

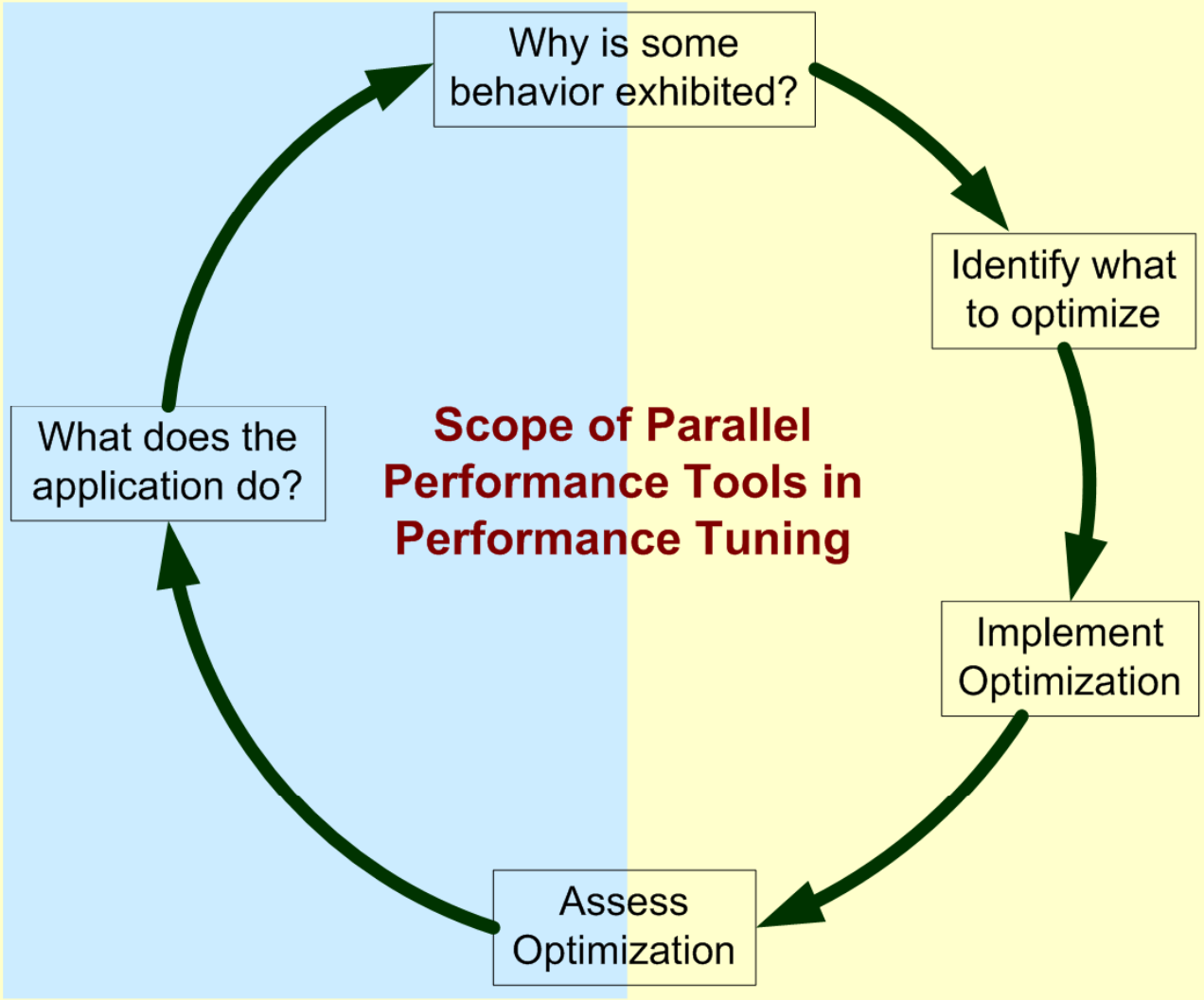
other functions

NFS Server Test Results – Application Tracing

Timeline for a Poor Performing Execution

Zoomed-In Portion of Timeline

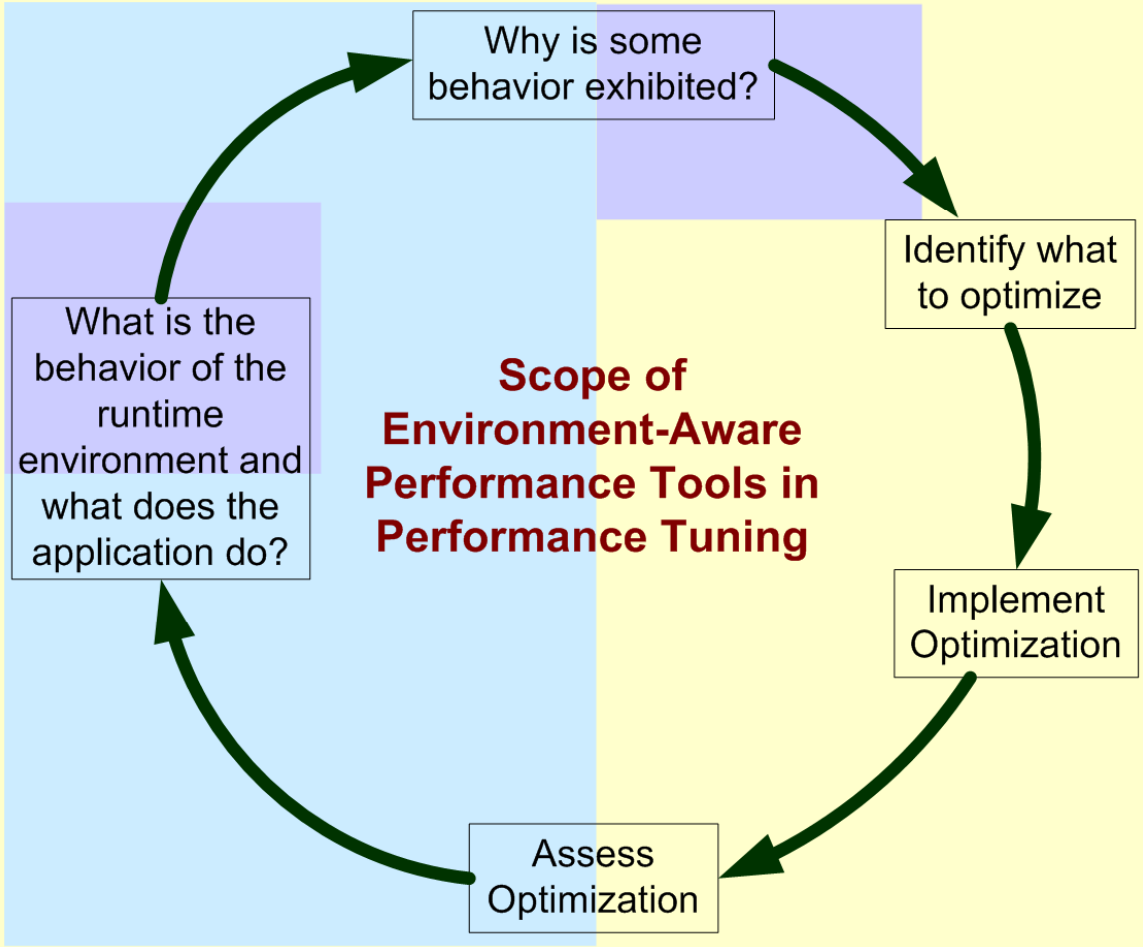
Scope of Parallel Performance Tools



Our Approach

Environment-Aware Performance Analysis

Automated methods to diagnose performance problems that are caused by the runtime system.



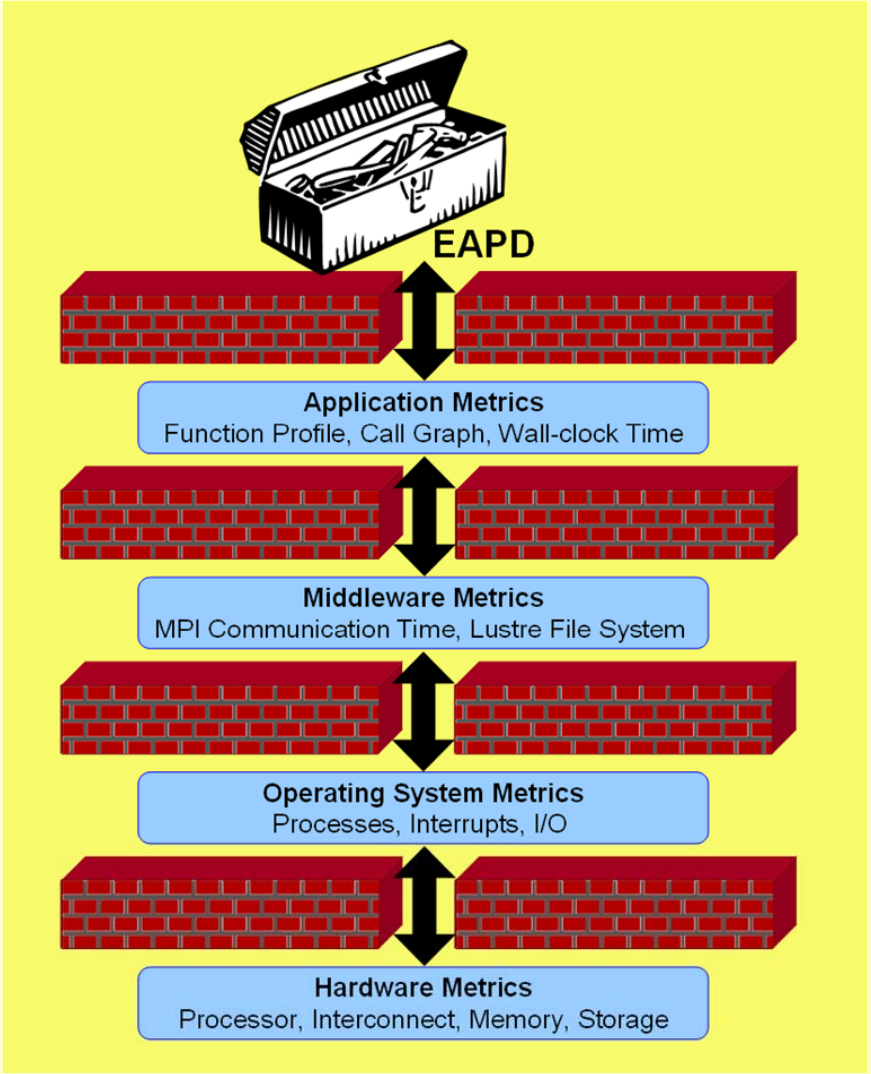
EAPD Challenges

Online EAPD

Low overhead measurement
Automated Diagnosis

Approach

Offline Diagnosis
Data Integration
- different tools
- different metrics
- PNNL collaboration



EPA reports energy used in U.S. for servers and data centers is significant.

- ~ 61 billion kilowatt-hours (kWh) in 2006
- 1.5% of total electricity consumption
- Total electricity cost of about $4.5 billion.
- Similar to the amount of electricity consumed by approximately 5.8 million average U.S. households (or about five percent of the total housing stock).
- Federal servers and data centers alone
  - ~ 6 billion kWh
  - 10% of electricity used for servers and data centers
  - Total electricity cost of about $450 million annually.

EPA Report to Congress on Server and Data Center Energy Efficiency **R**eleased On August 2, 2007 and in response to Public Law 109-431

The thrust of the Energy Smart Data Center at PNNL

Strategy

- Develop a testbed datacenter facility to promote energy efficiency in collaboration with other national laboratories, leaders of industry, and other energy-focused organizations.

Objectives

- Demonstrate and compare innovative cooling technologies
- Research potential savings in power conversion
- Partner with vendors and chip manufacturers to mature new technologies in a operational datacenter environment.
- Promote power aware computing

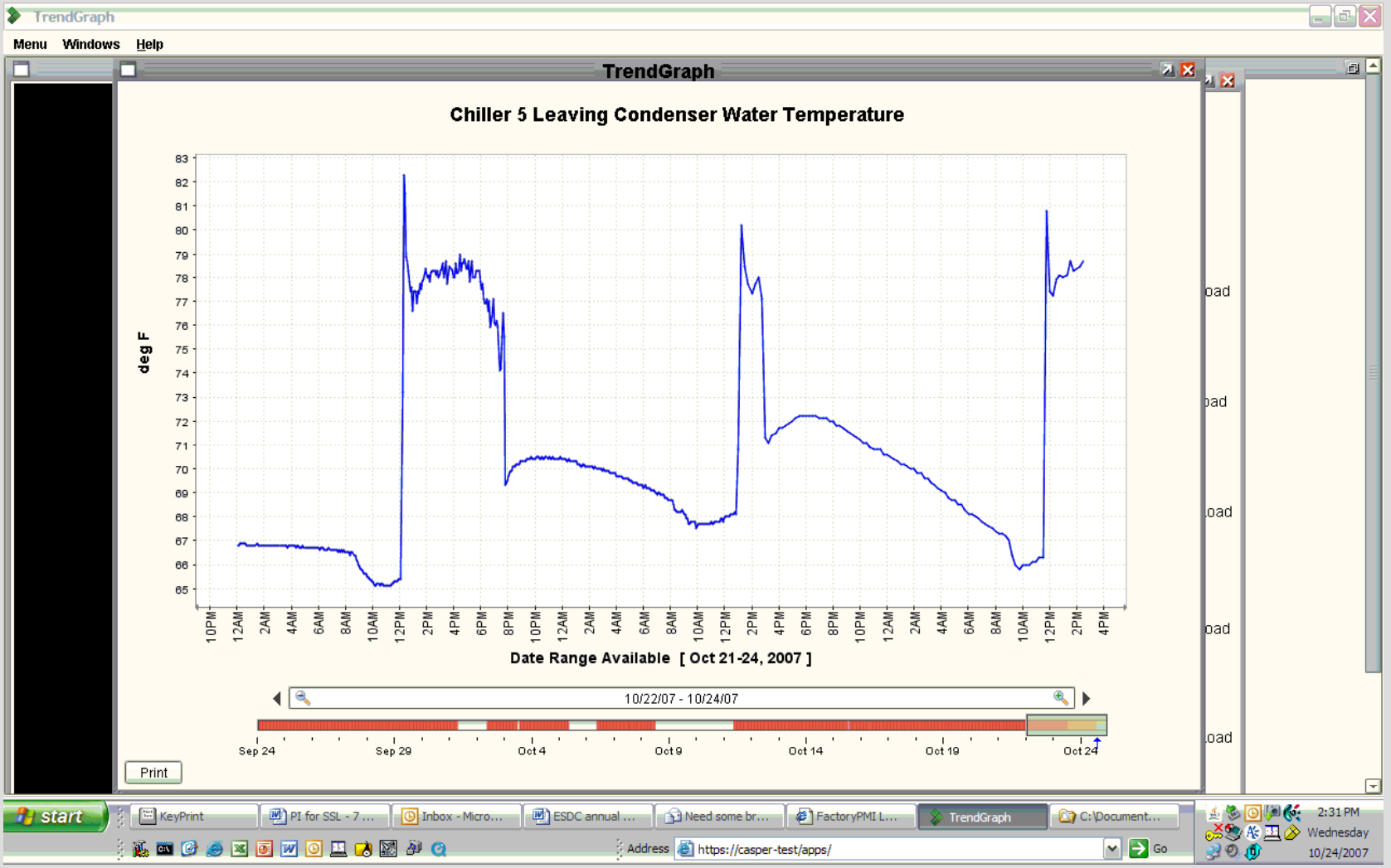
NW-ICE – First System To Be Assessed in the ESDC Testbed Facility

- 192 nodes, 2.3 GHz Intel (quad-core) Clovertown, 16 GB DDR2 FBDIMM memory,160 GB SATA disk, DDR Infiniband interconnect, dual GigE
- Five racks spraycooled
- Two racks air cooled
- Upcoming upgrades
  - global file system
- Data from
  - Building (MetaSys)
  - Sensors

Measurements at All Levels of the Infrastructure Hierarchy

- Server:
  - fan sensors, uP thermal diodes, server power, health (BMC/IPMI)
  - 3 servers/rack: temp at chipset, memory, in- egress (DAS)
- Rack:
  - air flow, power
  - TMU: PFC: flow, pressure, temp
- Computer room:
  - calorimetric zoning:
    - CW loop in- egress: flow, temp
    - Simulated CondensW loop in- egress: flow, temp
    - CW rack in- egress : flow, pressure, temp
  - Air (various points): temp
  - Air Handler: temp, RH, power
  - HVAC: temp, power
- Machine room:
  - CondensW in- egress: flow, temp, pump power, fan power, spray power
  - CW in- egress: flow, temp, mech. cooling power, primary-secondary pumps
- UPS:
  - power-in vs. power-out

Real-Time: Condenser Water Temperature Egress



Battelle

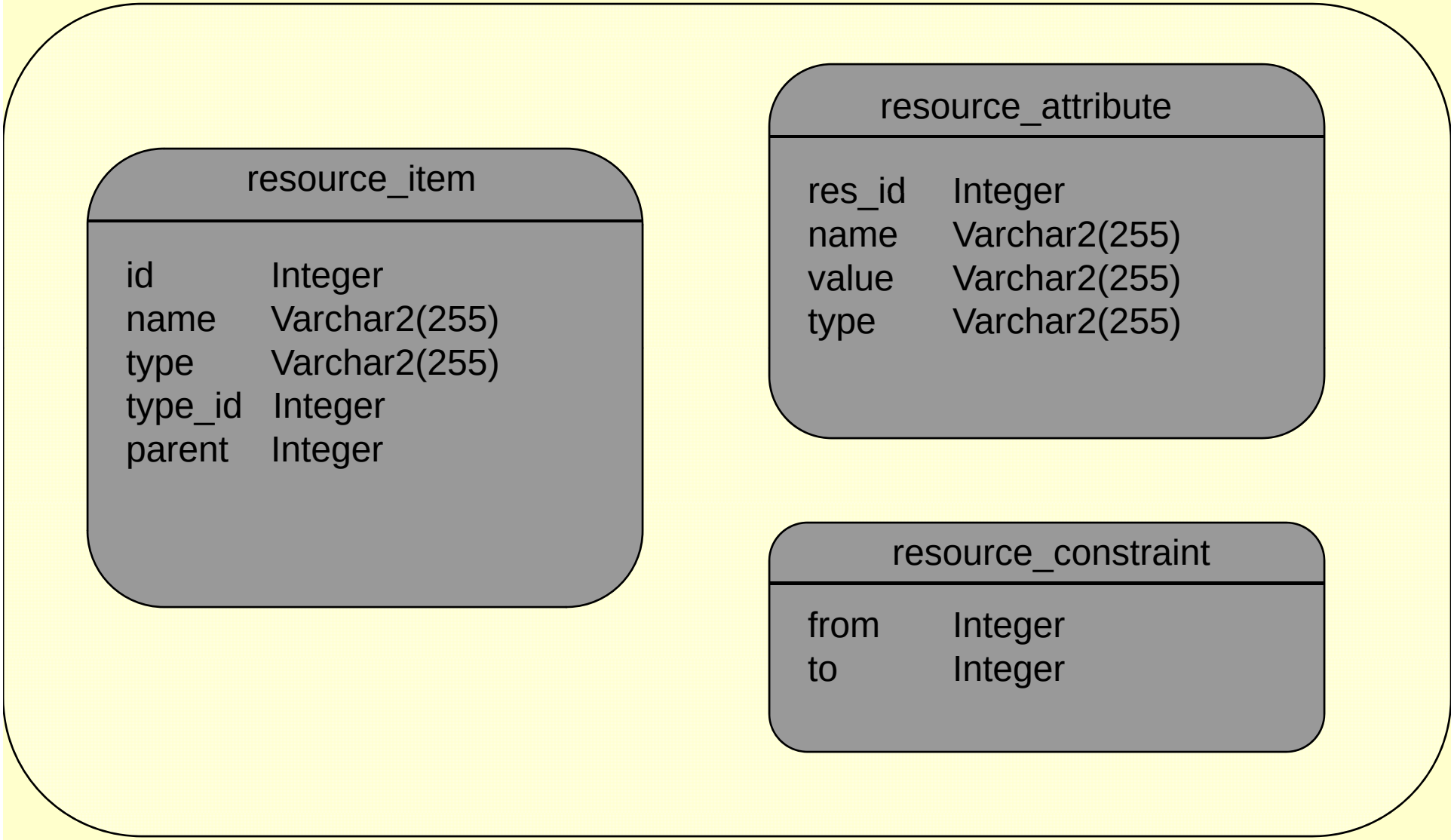
Pacific Northwest National Laboratory
U.S. Department of Energy

The PerfTrack Project

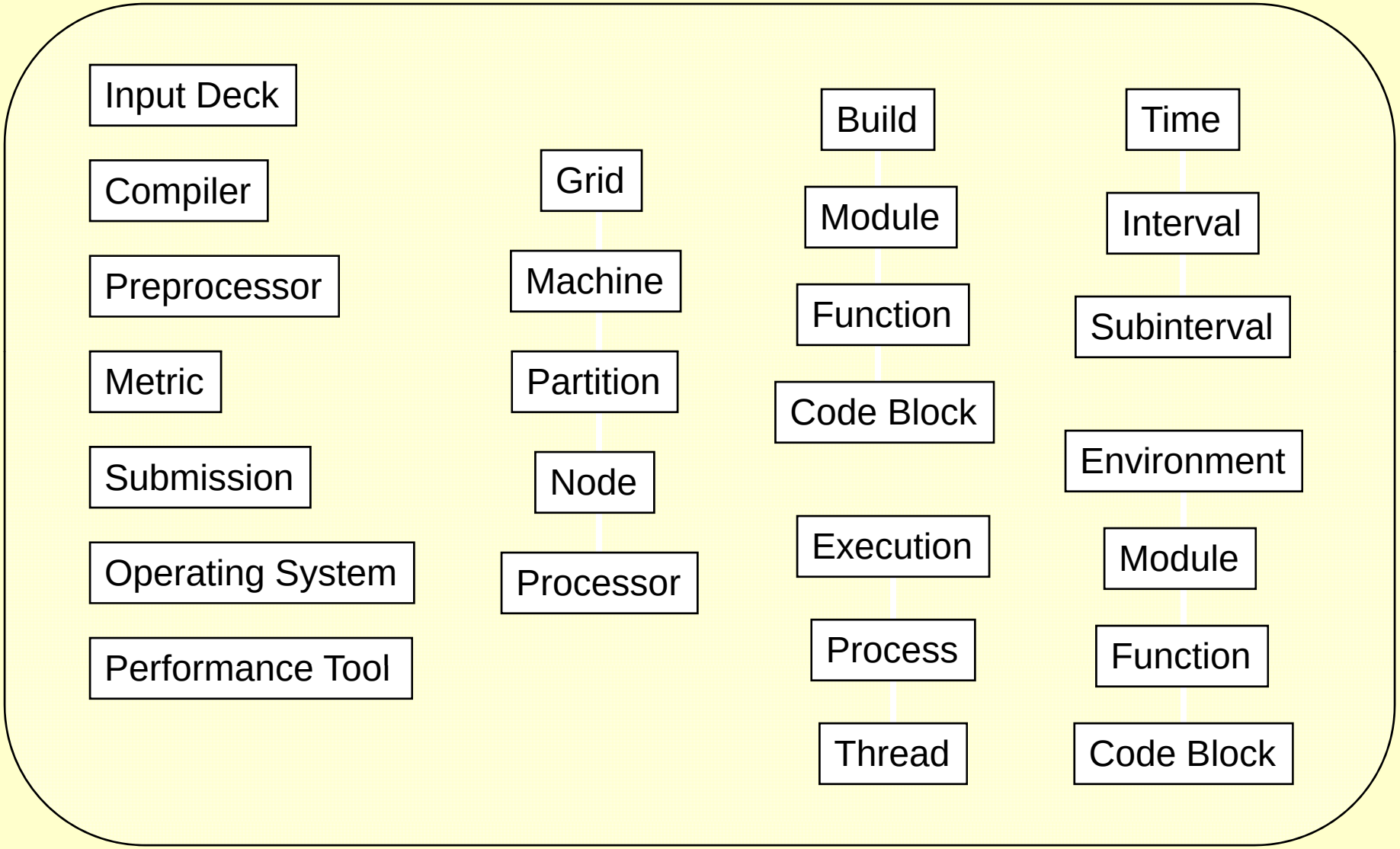
PerfTrack is a tool for storing, exploring, and analyzing application performance data

- Collect and store description of each build and run of an application
- Integrate DBMS into a performance analysis tool
- Store a wide variety of performance data
  - Data from different measurement tools
    - Tracing, DPCL, Paradyn, TAU, Vampir, Speedshop, HW counters, native application performance measurements, etc.
- GUI for data navigation and querying
- Shield tool user from DBMS internals

PerfTrack Design: Generic Database Schema



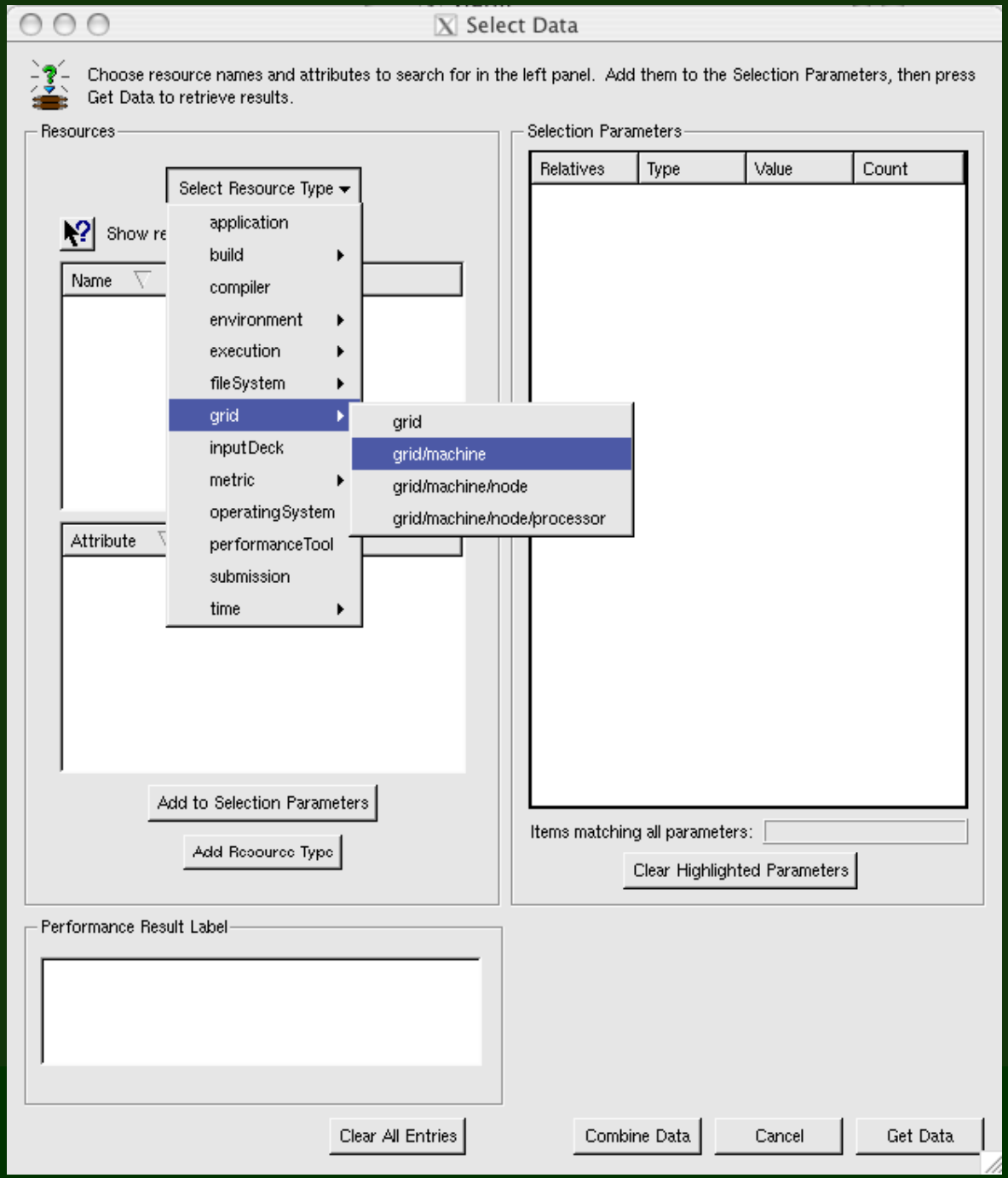
PerfTrack Design: Default Resource Types

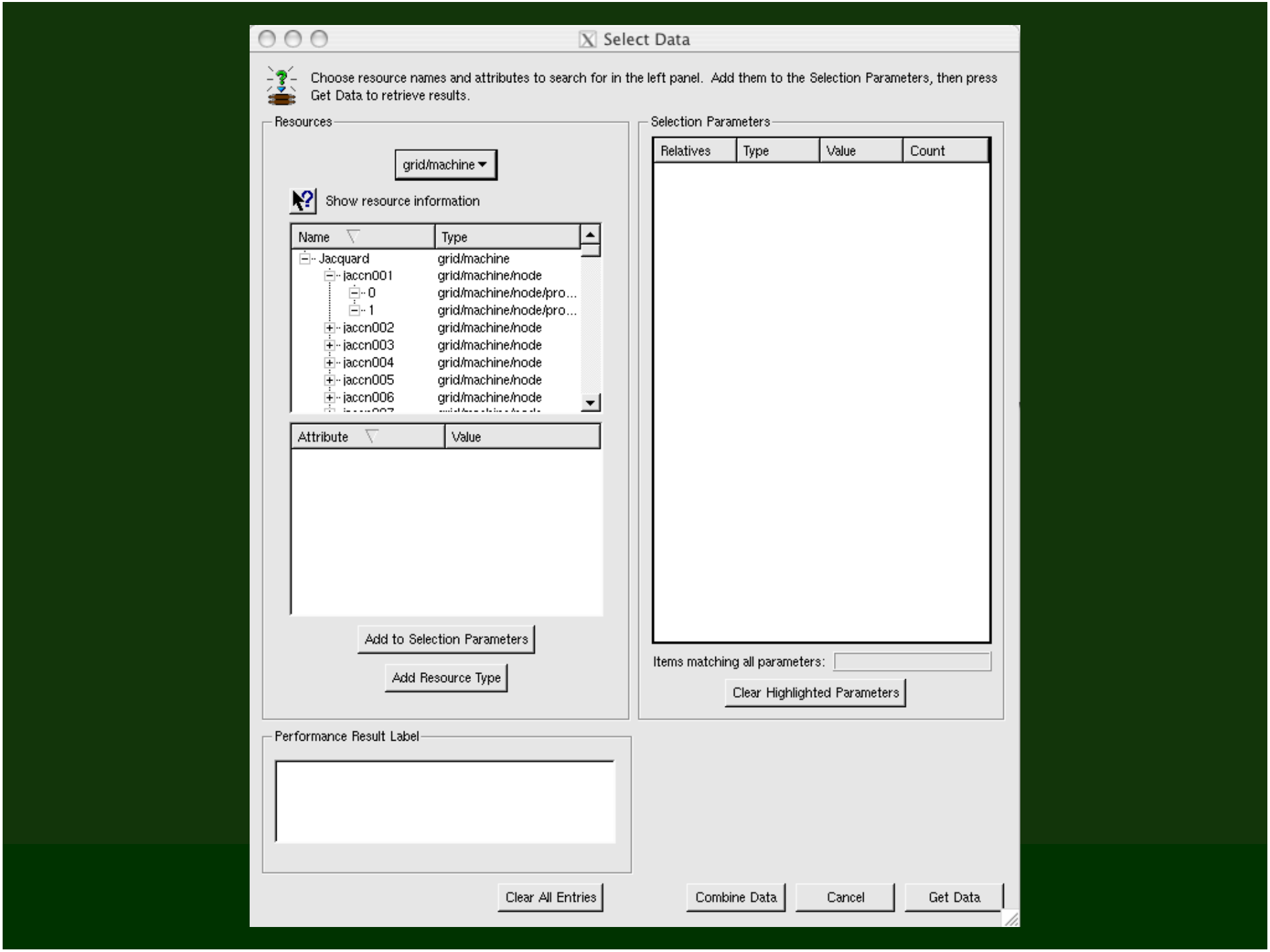


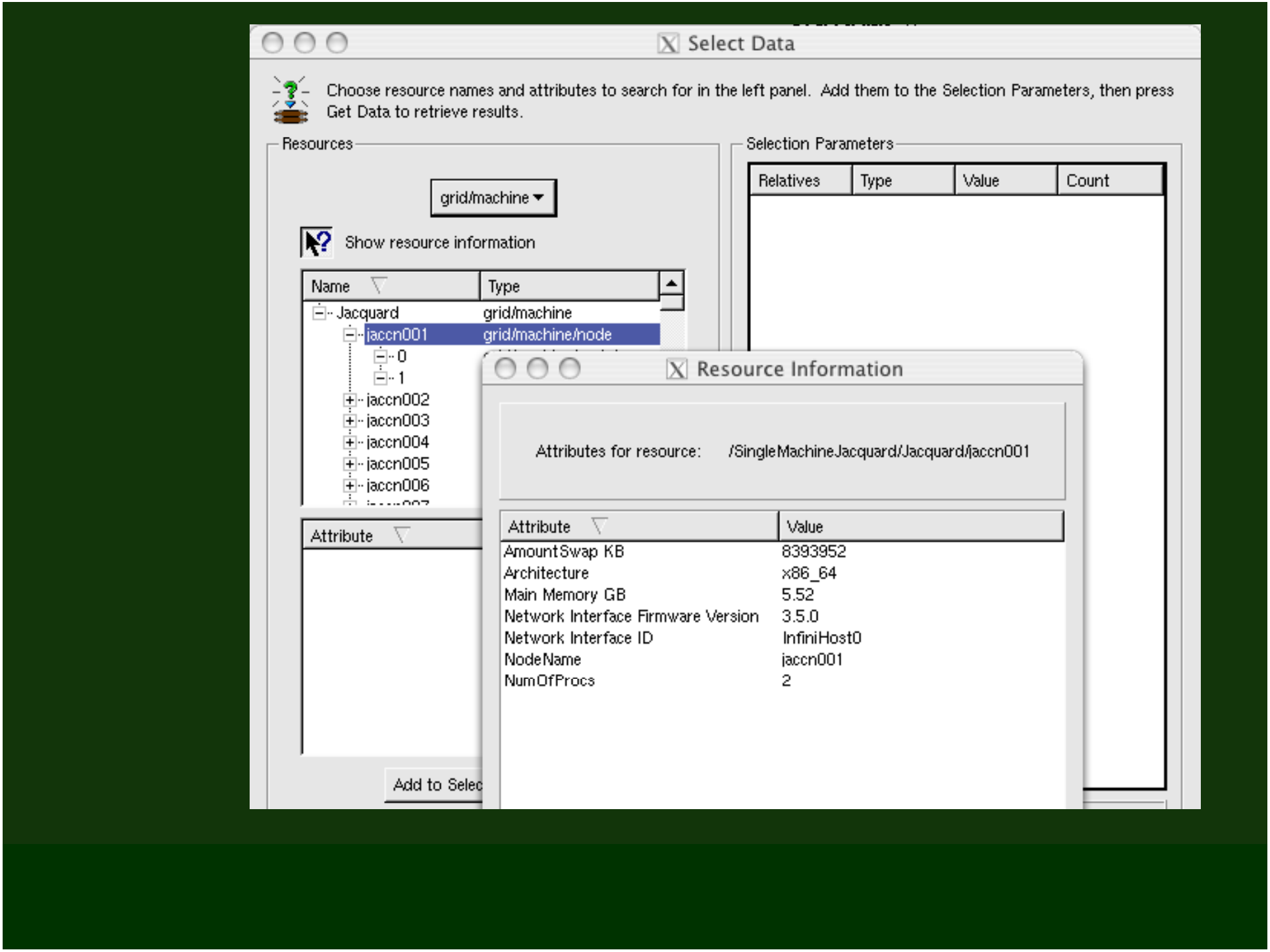
PerfTrack Design

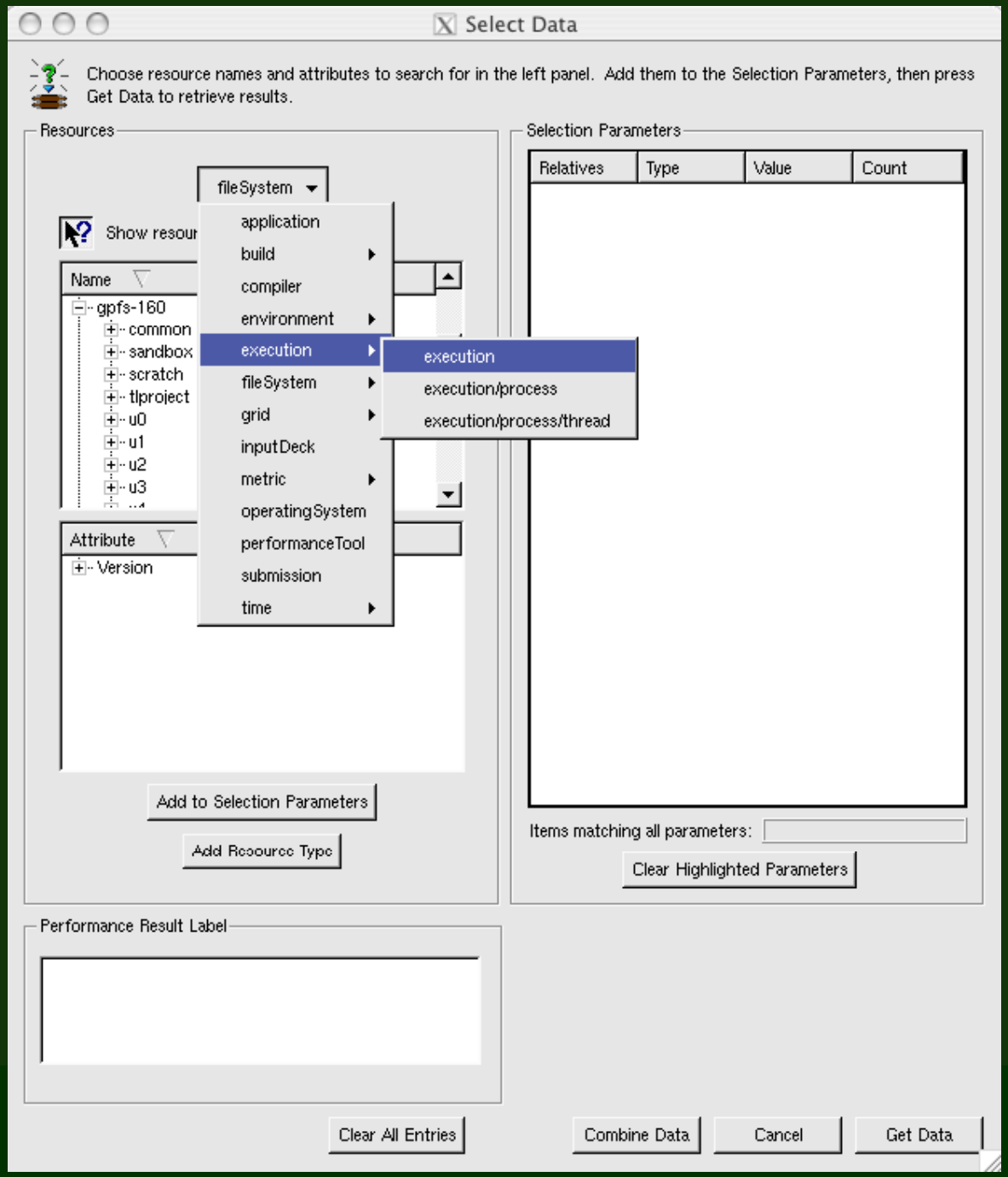
PerfTrack Data Format (PTdf):

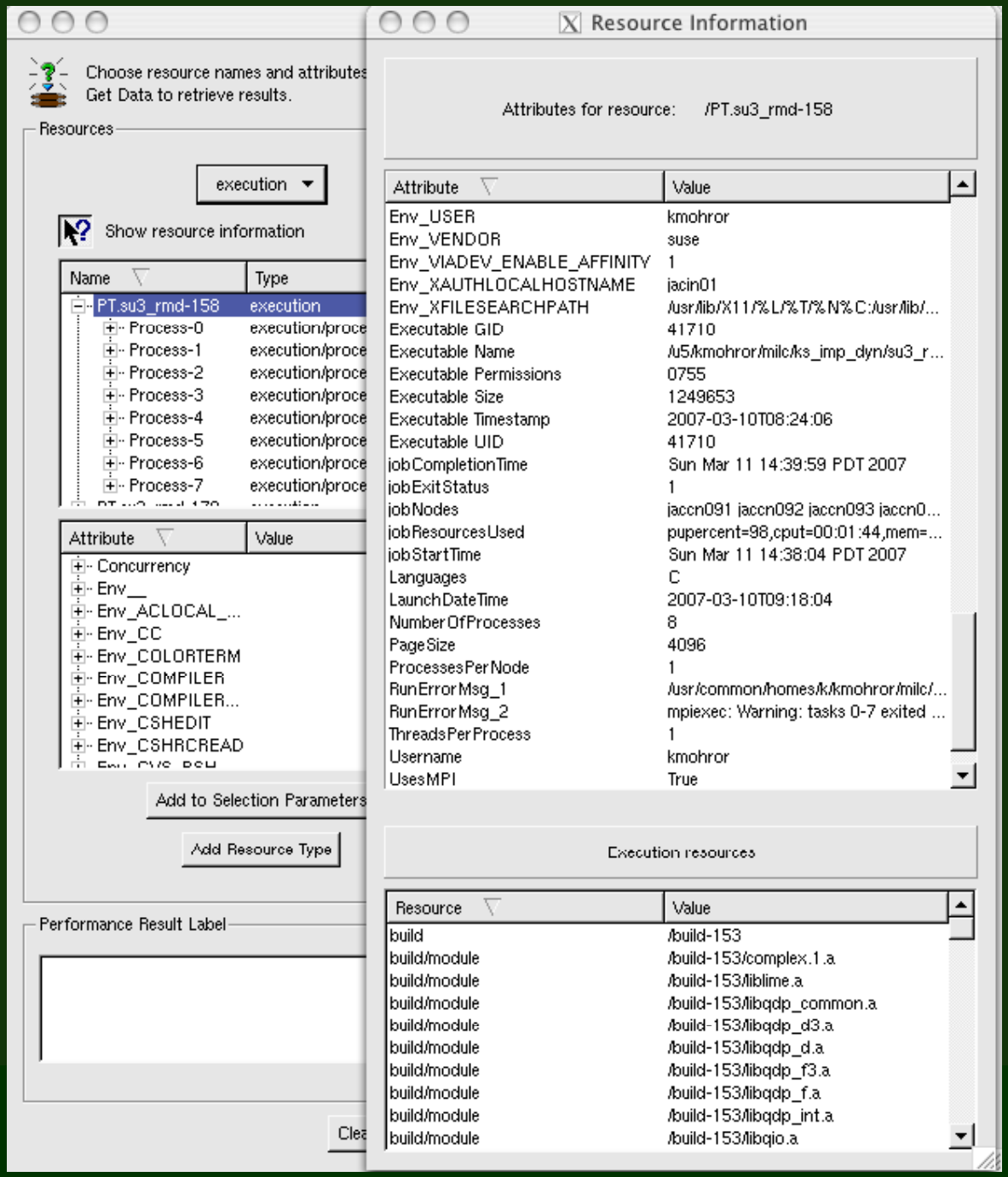
```
ResourceType resourceTypeName

Application appName

Execution execName appName

Resource resourceName resourceTypeName execName
Resource resourceName resourceTypeName

ResourceAttribute resourceName attributeName
attributeValue attributeType

ResourceConstraint resourceName1 resourceName2

PerfResult execName resourceSet perfToolName
metricName value units startTime endTime
```

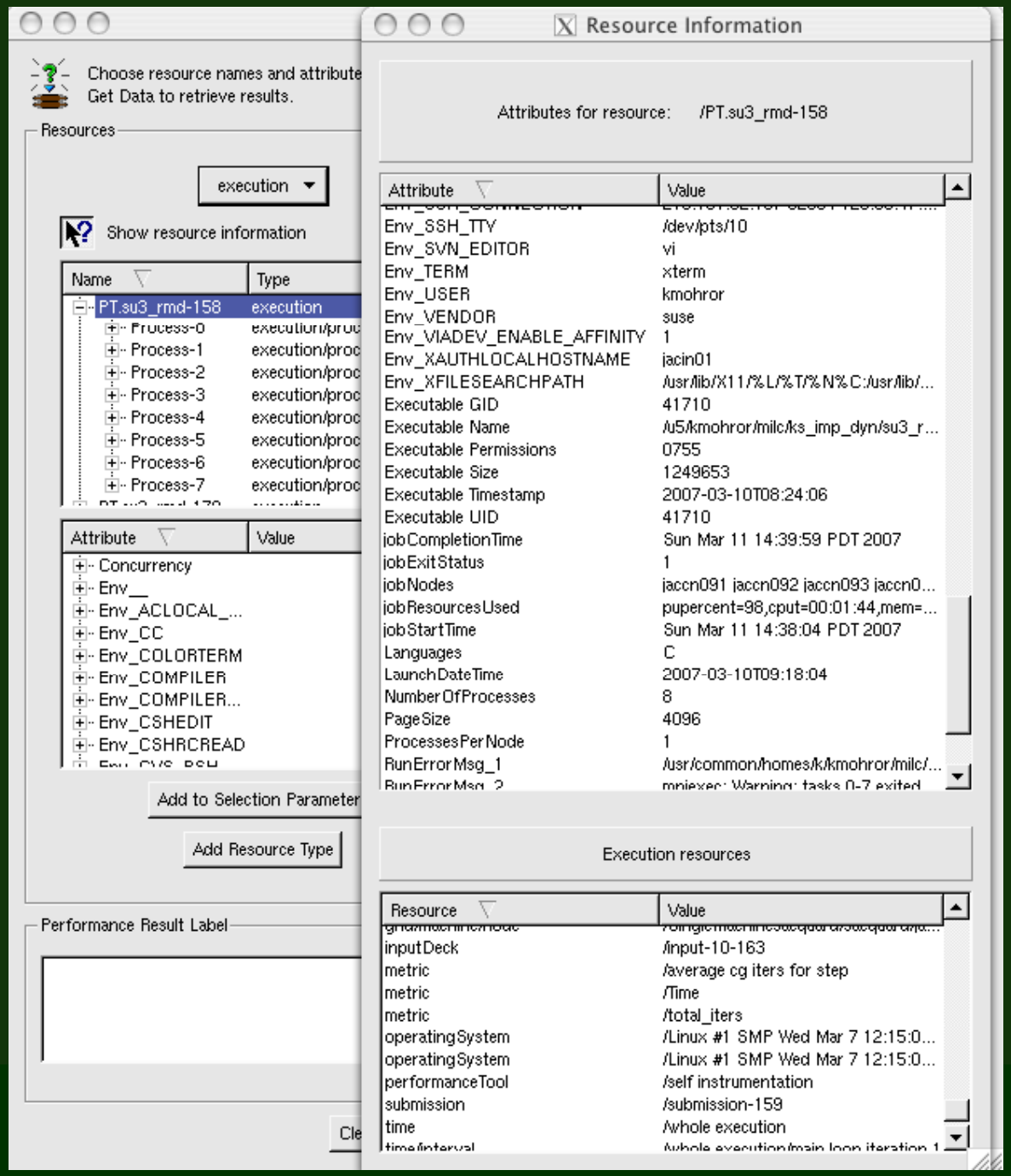

Select Data
Choose resource names and attributes to search for in the left panel. Add them to the Selection Parameters, then press Get Data to retrieve results.
Resources
Select Resource Type
application
build
compiler
environment
execution
fileSystem
grid
inputDeck
metric
operatingSystem
performanceTool
submission
time
grid
grid/machine
grid/machine/node
grid/machine/node/processor
Name
Attribute
Add to Selection Parameters
Add Resource Type
Selection Parameters
Relatives
Type
Value
Count
Items matching all parameters:
Clear Highlighted Parameters
Performance Result Label
Clear All Entries
Combine Data
Cancel
Get Data

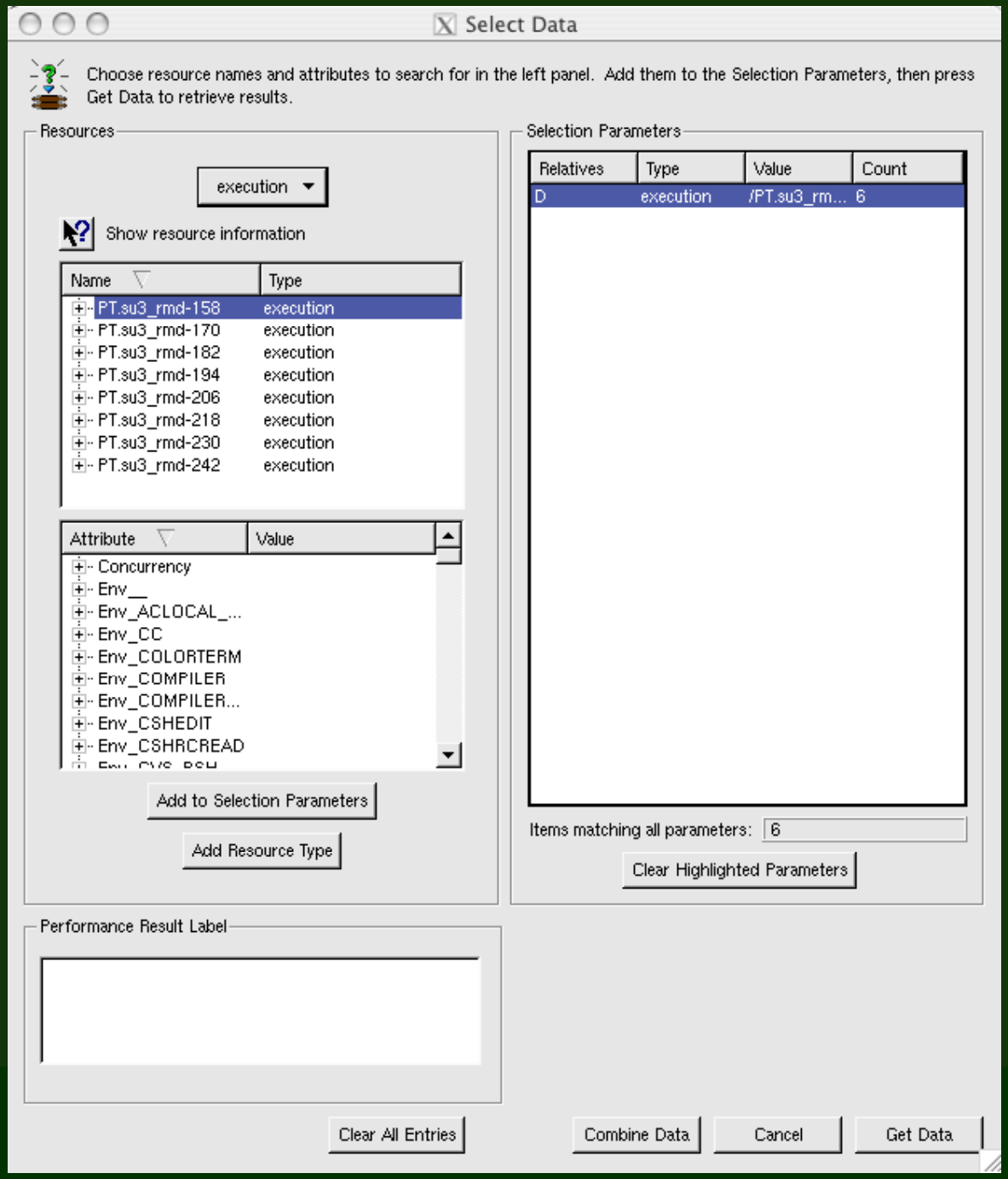

Select Data
Choose resource names and attributes to search for in the left panel. Add them to the Selection Parameters, then press Get Data to retrieve results.
Resources
grid/machine
Show resource information
Name
Type
Jacquard
grid/machine
iaccn001
grid/machine/node
0
grid/machine/node/pro...
1
grid/machine/node/pro...
iaccn002
grid/machine/node
iaccn003
grid/machine/node
iaccn004
grid/machine/node
iaccn005
grid/machine/node
iaccn006
grid/machine/node
Attribute
Value
Add to Selection Parameters
Add Resource Type
Selection Parameters
Relatives
Type
Value
Count
Items matching all parameters:
Clear Highlighted Parameters
Performance Result Label
Clear All Entries
Combine Data
Cancel
Get Data

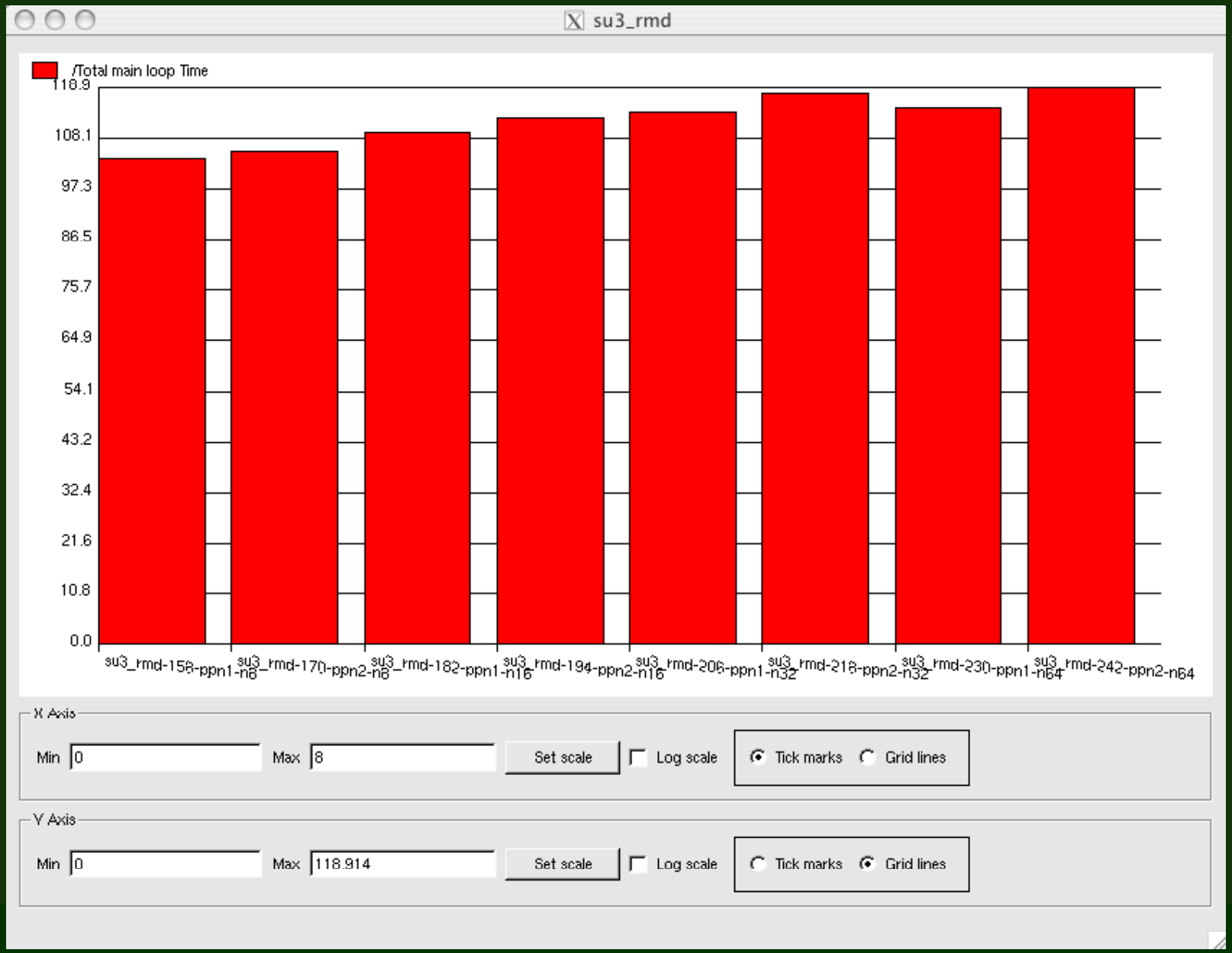

Select Data
Choose resource names and attributes to search for in the left panel. Add them to the Selection Parameters, then press Get Data to retrieve results.
Resources
grid/machine
Show resource information
Name
Type
Jacquard
grid/machine
jaccn001
grid/machine/node
0
1
jaccn002
jaccn003
jaccn004
jaccn005
jaccn006
Attribute
Add to Selec
Selection Parameters
Relatives
Type
Value
Count
Resource Information
Attributes for resource: /SingleMachineJacquard/Jacquard/jaccn001
Attribute
Value
AmountSwap KB
8393952
Architecture
x86_64
Main Memory GB
5.52
Network Interface Firmware Version
3.5.0
Network Interface ID
InfiniHost0
NodeName
jaccn001
NumOfProcs
2


Select Data
Choose resource names and attributes to search for in the left panel. Add them to the Selection Parameters, then press Get Data to retrieve results.
Resources
fileSystem
Show resour
Name
gpfs-160
common
sandbox
scratch
tlproject
u0
u1
u2
u3
Attribute
Version
application
build
compiler
environment
execution
fileSystem
grid
inputDeck
metric
operatingSystem
performanceTool
submission
time
execution
execution/process
execution/process/thread
Add to Selection Parameters
Add Resource Type
Selection Parameters
Relatives
Type
Value
Count
Items matching all parameters:
Clear Highlighted Parameters
Performance Result Label
Clear All Entries
Combine Data
Cancel
Get Data


Choose resource names and attributes
Get Data to retrieve results.
Resources
execution
Show resource information
Name
Type
PT.su3_rmd-158
execution
Process-0
Process-1
Process-2
Process-3
Process-4
Process-5
Process-6
Process-7
execution/proce
Attribute
Value
Concurrency
Env__
Env_ACLOCAL_...
Env_CC
Env_COLORTERM
Env_COMPILER
Env_COMPILER...
Env_CSHEDIT
Env_CSHRCREAD
Add to Selection Parameters
Add Resource Type
Performance Result Label
Resource Information
Attributes for resource: /PT.su3_rmd-158
Attribute
Value
Env_USER kmohror
Env_VENDOR suse
Env_VIADEV_ENABLE_AFFINITY 1
Env_XAUTHLOCALHOSTNAME jacin01
Env_XFILESEARCHPATH /usr/lib/X11/%L/%T/%N%C:/usr/lib/...
Executable GID 41710
Executable Name /u5/kmohror/milc/ks_imp_dyn/su3_r...
Executable Permissions 0755
Executable Size 1249653
Executable Timestamp 2007-03-10T08:24:06
Executable UID 41710
jobCompletionTime Sun Mar 11 14:39:59 PDT 2007
jobExitStatus 1
jobNodes jaccn091 jaccn092 jaccn093 jaccn0...
jobResourcesUsed pupercent=98,cput=00:01:44,mem=...
jobStartTime Sun Mar 11 14:38:04 PDT 2007
Languages C
LaunchDateTime 2007-03-10T09:18:04
NumberOfProcesses 8
PageSize 4096
ProcessesPerNode 1
RunErrorMsg_1 /usr/common/homes/k/kmohror/milc/...
RunErrorMsg_2 mpiexec: Warning: tasks 0-7 exited ...
ThreadsPerProcess 1
Username kmohror
UsesMPI True
Execution resources
Resource
Value
build /build-153
build/module /build-153/complex.1.a
build/module /build-153/liblime.a
build/module /build-153/libqdp_common.a
build/module /build-153/libqdp_d3.a
build/module /build-153/libqdp_d.a
build/module /build-153/libqdp_f3.a
build/module /build-153/libqdp_f.a
build/module /build-153/libqdp_int.a
build/module /build-153/libqio.a


Choose resource names and attribute
Get Data to retrieve results.
Resources
execution
Show resource information
Name
Type
PT.su3_rmd-158
execution
Process-0
Process-1
Process-2
Process-3
Process-4
Process-5
Process-6
Process-7
Attribute
Value
Concurrency
Env__
Env_ACLOCAL_...
Env_CC
Env_COLORTERM
Env_COMPILER
Env_COMPILER...
Env_CSHEDIT
Env_CSHRCREAD
Add to Selection Parameter
Add Resource Type
Performance Result Label
Resource Information
Attributes for resource: /PT.su3_rmd-158
Attribute
Value
Env_SSH_TTY
/dev/pts/10
Env_SVN_EDITOR
vi
Env_TERM
xterm
Env_USER
kmohror
Env_VENDOR
suse
Env_VIADEV_ENABLE_AFFINITY
1
Env_XAUTHLOCALHOSTNAME
jacin01
Env_XFILESEARCHPATH
/usr/lib/X11/%L/%T/%N%C:/usr/lib/...
Executable GID
41710
Executable Name
/u5/kmohror/milc/ks_imp_dyn/su3_r...
Executable Permissions
0755
Executable Size
1249653
Executable Timestamp
2007-03-10T08:24:06
Executable UID
41710
jobCompletionTime
Sun Mar 11 14:39:59 PDT 2007
jobExitStatus
1
jobNodes
jaccn091 jaccn092 jaccn093 jaccn0...
jobResourcesUsed
pupercent=98,cput=00:01:44,mem=...
jobStartTime
Sun Mar 11 14:38:04 PDT 2007
Languages
C
LaunchDateTime
2007-03-10T09:18:04
NumberOfProcesses
8
PageSize
4096
ProcessesPerNode
1
RunErrorMsg_1
/usr/common/homes/k/kmohror/milc/...
Execution resources
Resource
Value
inputDeck
/input-10-163
metric
/average cg iters for step
metric
/Time
metric
/total_iters
operatingSystem
/Linux #1 SMP Wed Mar 7 12:15:0...
operatingSystem
/Linux #1 SMP Wed Mar 7 12:15:0...
performanceTool
/self instrumentation
submission
/submission-159
time
/whole execution


Select Data
Choose resource names and attributes to search for in the left panel. Add them to the Selection Parameters, then press Get Data to retrieve results.
Resources
execution
Show resource information
Name
Type
PT.su3_rmd-158 execution
PT.su3_rmd-170 execution
PT.su3_rmd-182 execution
PT.su3_rmd-194 execution
PT.su3_rmd-206 execution
PT.su3_rmd-218 execution
PT.su3_rmd-230 execution
PT.su3_rmd-242 execution
Attribute
Value
Concurrency
Env__
Env_ACLOCAL_...
Env_CC
Env_COLORTERM
Env_COMPILER
Env_COMPILER...
Env_CSHEDIT
Env_CSHRCREAD
Add to Selection Parameters
Add Resource Type
Selection Parameters
Relatives
Type
Value
Count
D execution /PT.su3_rm... 6
Items matching all parameters: 6
Clear Highlighted Parameters
Performance Result Label
Clear All Entries
Combine Data
Cancel
Get Data


su3_rmd
/Total main loop Time
118.9
108.1
97.3
86.5
75.7
64.9
54.1
43.2
32.4
21.6
10.8
0.0
su3_rmd-158-ppn1-n8
su3_rmd-170-ppn2-n8
su3_rmd-182-ppn1-n16
su3_rmd-194-ppn2-n16
su3_rmd-206-ppn1-n32
su3_rmd-218-ppn2-n32
su3_rmd-230-ppn1-n64
su3_rmd-242-ppn2-n64
X Axis
Min 0
Max 8
Set scale
Log scale
Tick marks
Grid lines
Y Axis
Min 0
Max 118.914
Set scale
Log scale
Tick marks
Grid lines

Integrating Data Center Results
New Resource Types

Default Resource Hierarchy

grid
grid|machine
grid|machine|partition
grid|machine|partition|node *
grid|machine|partition|node|processor

New Node Attributes

Number of Sockets

Number of Cores

Additional Resource Types

grid|machine|partition|node|dimm
grid|machine|partition|node|ASIC

datacenter
datacenter|coolingtower
datacenter|chiller
datacenter|heatrecovery
datacenter|rack *
datacenter|rack|powerunit
datacenter|rack|TMU
datacenter|airconditioner
datacenter|pump

Integrating Data Center Results Measurement Data

New metrics

Cooling Tower: EFan KW, WFan KW, EFan Speed, etc

Nodes: CPU Temp, Dimm Temp

Racks: Supply temp, Return temp

Removed an exception to the Generic Schema: "Execution"

- PerfResult execName focus perfTool metricName value units sTime eTime
- Result resourceName focus perfTool metricName value units sTime eTime

Correlating room data to applications

- Time Scales
- Link each application run to its nodes

Summary

Correct diagnosis of performance problems is challenging

- Requires methods that combine application and system level metrics
- The search space of possible diagnoses is too large
  - Not enough time to apply manual detection methods
  - Not enough resource allocation for online methods
  - --> automated and online methods that reduce the search space and order the traversal of the search space

EAPD model for automated diagnosis

- Combines environment, system, and application level metrics

PerfTrack performance data store

Integration of data from different layers

PSU High Performance Computing Lab

http://www.cs.pdx.edu/~karavan/hpcLab.html

See our Short Demo at SC08

PerfTrack

Collaborator: Dr. John May (Lawrence Livermore National Laboratory)

Contributing PSU students: Kathryn Mohror, Rashawn Knapp, Brian Pugh, Dylan Enloe, Aaron Amauba

Contributing H.S. students: Travis Chapman (OSU), Thomas Conerly (CMU), Abraham Neben (Northwestern)

Environment-Aware Performance Diagnosis

Collaborators: Dr. Douglas Pase (IBM), Dr. Andres Marquez (PNNL)

Contributing PSU students: Rashawn Knapp, Agniv Adhikari, Mike Smith, Dave Revell, Dylan Enloe

This research supported in part by: the PSU Center for Sustainable Processes and Practices, PNNL/Battelle, UC/LLNL and the DOE Office of Science.