Scalable, Automated Parallel Performance Analysis with TAU, PerfDMF and PerfExplorer

Kevin A. Huck, Allen D. Malony, Sameer Shende, Alan Morris
khuck, malony, sameer, amorris@cs.uoregon.edu
http://www.cs.uoregon.edu/research/tau
University of Oregon
Department of Computer and Information Science
Performance Research Laboratory
Performance Research Lab

- Prof. Allen D. Malony
- Sameer Shende
- Alan Morris
- Wyatt Spear
- Scott Biersdorf
- Aroon Nataraj
- Kevin A. Huck
- http://www.cs.uoregon.edu/research/tau/
Overview

• TAU Introduction (brief)
• PerfDMF Introduction
• PerfExplorer Introduction
• PerfExplorer Ongoing Analysis Examples
• Summary
TAU Performance System

- Tuning and Analysis Utilities (15+ year project effort)
- Performance system framework for HPC systems
  - Integrated, scalable, flexible and parallel
- Targets a general complex system computation model
  - Entities: nodes / contexts / threads
- Multi-level: system / software / parallelism
  - Measurement and analysis abstraction
- Integrated toolkit for performance problem solving
  - Instrumentation, measurement, analysis and visualization
- Portable performance profiling and tracing utility
- Performance data management and data mining
- **Partners**: Forshungszentrum Jülich, LLNL, ANL, LANL
PerfDMF

- **Performance Data Management Framework**
- Originally designed to address critical TAU requirements
- Broader goal is to provide an open, flexible framework to support common data management tasks
- Extensible toolkit to promote integration and reuse across available performance tools

- Supported profile formats:
  - TAU, CUBE, Dynaprof, HPCToolkit (Rice), HPC Toolkit (IBM), gprof, mpiP, psrun (PerfSuite), OpenSpeedShop (in development), application

- Supported DBMS:
  - PostgreSQL, MySQL, Oracle, DB2, Derby/Cloudscape
PerfDMF Architecture

Display & Analysis Tools

ParaProf  ...  PerfExplorer

Parsed / Importers

metadata
Profiles

PerfDMF Query & Analysis Toolkit

Exporters

XML
Formatted Profiles

PostgreSQL, MySQL,
DB2, Oracle, Derby

RDBMS

TAU, CUBE, Dynaprof,
mpiP, HPCToolkit,
IBM HPC Toolkit,
Dynaprof, psrun
Recent PerfDMF Development

- Integration of XML metadata for each profile
  - Common profile attributes
  - Thread / process specific profile attributes
  - Automatic collection of runtime information
  - Any other data the user wants to collect can be added
    - build information
    - job submission information
- Two methods for acquiring metadata:
  - TAU_METADATAD() call from application
  - Optional XML file added when saving profile to PerfDMF
- TAU Metadata XML schema is simple, easy to generate from scripting tools (no XML libraries required)
PerfExplorer

- Performance knowledge discovery framework
- Data mining analysis applied to parallel performance data
  - comparisons, clustering, correlation, dimension reduction, ...
- Uses the existing TAU infrastructure
  - TAU & other performance profiles, PerfDMF
- Technology Integration
  - Java API and toolkit for portability
  - R-project / Omegahat statistical analysis
  - Weka data mining package
  - JFreeChart for visualization, output (EPS, SVG, PNG)
PerfExplorer Design

GUI

Scripting Interface

PerfExplorer Component Interfaces

Analysis Components

Data Mining

Inference Engine

Data Components

PerfDMF

Expert Rules

Analysis Results

Provenance

Performance Data

Metadata

R

Weka

ParCo 2007, RWTH Aachen / Forschungszentrum Jülich

September 5, 2007
PerfExplorer Analysis Methods

- Data summaries, distributions, scatterplots
- Clustering
  - k-means
  - hierarchical
- Correlation analysis
- Dimension reduction
  - PCA
  - Random Linear Projection
  - Thresholds
- Comparative analysis
- Data management views
Relative Comparisons

- Total execution time
- Timesteps per second
- Relative efficiency / speedup for total / per event
- Group fraction of total
- Runtime breakdown
- Phase comparisons
Distribution Visualization

Indicates min, Q1, mean, median, Q3, max, outliers

Visualizes Intra-Quartile Range (median 50%)

Observation of multi-modal distributions (clusters)
Cluster Analysis

![Cluster Analysis Graph](image-url)
Correlation Analysis

An “interesting” relationship?

Data: FLASH on BGL (LLNL)
4-D Visualization

4 “significant” events are selected

clusters and correlations are visible
New PerfExplorer Features:

- Additional sophisticated metadata support
- Scripting using Jython*
- Inference engine to “infer” causes of performance phenomena from expert rules
- Redesigned component-based analysis
- Persistence of intermediate results
- Provenance

* or any other scripting language
Example: Regression Testing
Example: Parametric Study
PerfExplorer Example: S3D

- Compressible Navier-Stokes solver coupled with an integrator for detailed chemistry (CHEMKIN-compatible)
- Models turbulent reacting flow in combustion science
- MPI-based parallel computing implementation

S3D Scaling: Efficiency

Relative Efficiency of S3D, Including Serial Execution

Up to 47% loss of efficiency
baseline: serial execution
S3D Scaling: Parallel Efficiency

Relative Efficiency of S3D, Parallel Only

up to 30% loss of efficiency
(baseline: 8 processors)
S3D Scaling: Correlation

Correlation Results: S3D (Jaguar, ORNL):Harness Scaling Study: GET_TIME_OF_DAY

- MPI_WAIT

Legend:
- Loop: CHEMINFO_M::REACTION_RATE_BOUNDS [([cheminfo_m.pp.f90] [374,3]--[386,7]), r = 0.09]
- Loop: DERIVATIVE_X_CALC [([derivative_x.pp.f90] [432,10]--[441,15]), r = -0.10]
- Loop: DERIVATIVE_Y_CALC [([derivative_y.pp.f90] [431,10]--[440,15]), r = -0.16]
- Loop: DERIVATIVE_Z_CALC [([derivative_z.pp.f90] [435,10]--[444,15]), r = -0.02]
- Loop: INTEGRATE [([integrateerk.pp.f90] [73,3]--[93,13]), r = 0.37]
- Loop: RHSF [([rhsf.pp.f90] [209,3]--[211,7]), r = 0.01]
- Loop: RHSF [([rhsf.pp.f90] [515,3]--[535,16]), r = -0.13]
- Loop: RHSF [([rhsf.pp.f90] [537,3]--[543,16]), r = -0.30]
- Loop: THERMINFO_M::CALC_INV_AVG_MOL_WT [([therminfo_m.pp.f90] [127,5]--[129,9]), r = 0.34]
- Loop: THERMINFO_M::CALC_SPECEVENT_ALLPTS [([therminfo_m.pp.f90] [506,3]--[512,8]), r = -0.28]
- Loop: THERMINFO_M::CALC/temp [([therminfo_m.pp.f90] [175,5]--[216,9]), r = -0.68]
- Loop: TRANSPORT::COEFFICIENTS [([mixavg_transport_mappppp.f90] [492,5]--[520,9]), r = 0.69]
- Loop: TRANSPORT::COMPUTEHEATFLUX [([mixavg_transport_mappppp.f90] [782,5]--[790,19]), r = 0.19]
- Loop: TRANSPORT::COMPUTE_SPECIES_DIFFFLUX [([mixavg_transport_mappppp.f90] [630,5]--[656,19]), r = 0.28]
- Loop: VARIABLES::GET_MASS_FRAC [([variables_m.pp.f90] [96,3]--[99,7]), r = 0.34]
- MPI_Comm_compare(), r = 0.90
- WRITE_SAVEFILE [([io.pp.f90] [240,14]), r = 0.77]
- TOTAL
Most nodes are waiting on a few “slower” nodes...

6400 processes total

ranks 0-113

ranks 3200-3313
Cluster explanation...

Metadata identifies node names

- Ranks 0-113 lie on processors 3406-3551
- Ranks 3200-3313 are also on 3406-3551
- 3406-3551 are on the XT3 partition
- XT3 has slower DDR-400 memory (5986 MB/s)
- XT3 has a slower SS1 (1109 MB/s) interconnect
- XT4 partition has faster DDR2-667 memory modules (7147 MB/s) and faster Seastar2 (SS2) (2022 MB/s) interconnect
- Running on XT4 yields 12% improvement
- If XT3/XT4, load balancing will be required
S3D: “Improved” Result

390 sec.  MPI_Wait

104 sec.
PerfExplorer example: GTC

- GTC: Gyrokinetic Toroidal Code
- Particle-in-cell physics simulation
- CHARGEI: particles apply their charge to the grid cells
- PUSHI: cells update particle locations
- MPI-based parallel implementation

Figure: The (turbulent) electrostatic potential from a GYRO simulation of plasma microturbulence in the DIII-D tokamak.
GTC: Dynamic Phases

- L1 hit rate ~0.994
- L2 hit rate ~0.92
- MFLOPS ~1120

Source: GTC on Cray XT4
GTC: Iteration $i \Rightarrow$ PUSHI

“sawtooth” caused by diagnostic step every fourth iteration

L1 hit rate

.994

.988

L2 hit rate

.95

.87

MFLOPS

1700

1450

Source: GTC on Cray XT4
GTC: Iteration $i \mapsto$ CHARGEI

- **L1 hit rate**
  - Source: GTC on Cray XT4

- **L2 hit rate**
  - "sawtooth" gone

- **MFLOPS**
  - 710
  - 600
GTC: after 5,000 Iterations

GTC Phase Breakdown: GET_TIME_OF.DAY

Source: GTC on BG/L

it only continues to get worse...
GTC Phases: Summary

• CHARGEI, PUSHI events have good spatial locality of particles, but bad temporal locality for cells

• After 100 iterations, each iteration on Cray XT3/XT4 takes ~1 second longer (~12%)

• Full simulation runs for 10,000 timesteps (potential improvement of 2+ hours from 20 hour execution)

• PUSH event calls a diagnostic routine every $X$ timesteps (input variable)

• Analysis is ongoing...
Conclusion

- **TAU:**
  - Portable, configurable, complete instrumentation and measurement of parallel profiles and traces

- **PerfDMF:**
  - Profile management, query and analysis API
  - Supports most commonly used profile formats (if not, we can add it)

- **PerfExplorer:**
  - Robust profile parametric study support
  - In-depth analysis of large scale profiles
Acknowledgments

- US Department of Energy (DOE)
- Office of Science
  - MICS, Argonne National Lab
- ASC/NNSA
  - University of Utah ASC/NNSA Level 1
  - ASC/NNSA, Lawrence Livermore National Lab
- US Department of Defense (DoD)
- NSF Software and Tools for High-End Computing
- Forschungszentrum Jülich
- TU Dresden
- Los Alamos National Laboratory
- ParaTools, Inc.
- SciDAC, PERI, ORNL, NERSC, RENCI