Understanding applications using the BSC performance tools

Judit Gimenez (judit@bsc.es), Harald Servat (harald@bsc.es)
Humans are visual creatures

- Films or books?
  - Two hours vs. days (months)

- Memorizing a deck of playing cards
  - Each card translated to an image (person, action, location)

- Our brain loves pattern recognition
  - What do you see on the pictures?

PROCESS

STORE

IDENTIFY
Our Tools

- Since 1991
- Based on traces
- Open Source
  - http://www.bsc.es/paraver
- Core tools:
  - Paraver (paramedir) – offline trace analysis
  - Dimemas – message passing simulator
  - Extrae – instrumentation
- Focus
  - Detail, variability, flexibility
  - Behavioral structure vs. syntactic structure
  - Intelligence: Performance Analytics
Paraver
Paraver – Performance data browser

- Raw data

- Trace visualization/analysis + trace manipulation

- Timelines

- 2/3D tables (Statistics)

- Goal = Flexibility
  - No semantics
  - Programmable

- Comparative analyses
  - Multiple traces
  - Synchronize scales
Timelines

Each window displays one view
- **Piecewise constant** function of time

Types of functions
- Categorical
  - State, user function, outlined routine
- Logical
  - In specific user function, In MPI call, In long MPI call
- Numerical
  - IPC, L2 miss ratio, Duration of MPI call, duration of computation burst

\[ s(t) = S_i, i \in [t_i, t_{i+1}) \]

\[ S_i \in [0, n] \subset N, \quad n < \]

\[ S_i \in \{0, 1\} \]

\[ S_i \in R \]
Tables: Profiles, histograms, correlations

From timelines to tables

MPI calls profile

Useful Duration
Analyzing variability through histograms and timelines

Useful Duration

IPC

Instructions

L2 miss ratio
By the way: six months later ....
From tables to timelines

Where in the timeline do the values in certain table columns appear?

ie. want to see the time distribution of a given routine?
Variability … is everywhere

- CESM: 16 processes, 2 simulated days
- Histogram useful computation duration shows high variability
- How is it distributed?

- Dynamic imbalance
  - In space and time
  - Day and night.
  - Season? 😊
Trace manipulation

Data handling/summarization capability

– Filtering
  • Subset of records in original trace
  • By duration, type, value,…
  • Filtered trace IS a paraver trace and can be analysed with the same cfgs (as long as needed data kept)

– Cutting
  • All records in a given time interval
  • Only some processes

– Software counters
  • Summarized values computed from those in the original trace emitted as new event types
  • #MPI calls, total hardware count,…
Dimemas: Coarse grain, Trace driven simulation

Simulation: Highly non linear model

- Linear components
  - Point to point communication
  - Sequential processor performance
    - Global CPU speed
    - Per block/subroutine

- Non linear components
  - Synchronization semantics
    - Blocking receives
    - Rendezvous
    - Eager limit
  - Resource contention
    - CPU
    - Communication subsystem
      » links (half/full duplex), busses

\[ T = \frac{\text{MessageSize}}{\text{BW}} + L \]
Dimemas: Type of analysis

- **Parametric sweeps**
  - On abstract architectures
  - On application computational regions

- **What if analysis**
  - Ideal machine (instantaneous network)
  - Estimating impact of ports to MPI+OpenMP/CUDA/…
  - Should I use asynchronous communications?
  - Are all parts of an app. equally sensitive to network?

- **MPI sanity check**
  - Modeling nominal

Detailed feedback on simulation (trace)
Paraver & Dimemas: Ecosystem and Integration

- Paraver – Dimemas tandem
  - Analysis and prediction
  - What-if from selected time window

- Analytics
  - Clustering to model speeding or balancing regions

- Scalability model
  - Separate transfer and serialization efficiencies

- Coupled simulation
  - Multiscale simulation
  - Combine with detailed network / CPU simulators

Network sensitivity

MPIRE 32 tasks, no network contention

- $L = 25\mu s$ – $BW = 100\text{MB/s}$
- $L = 1000\mu s$ – $BW = 100\text{MB/s}$
- $L = 25\mu s$ – $BW = 10\text{MB/s}$

All windows same scale
Ideal machine

The impossible machine: \( BW = \infty, \quad L = 0 \)

- Load balance problems?
- Dependence problems?

GADGET @ Nehalem cluster
256 processes

Real run

Ideal network

Impact on practical machines?
Impact of architectural parameters

Ideal speeding up ALL the computation bursts by the CPU ratio factor

- The more processes the less speedup (higher impact of bandwidth limitations)!!
Models
\[ \eta_{\parallel} = LB \times \text{CommEff} \]

Directly from real execution metrics

\[ CommEff = \max(\text{eff}_i) \]

\[ LB = \frac{\sum_{i=1}^{P} \text{eff}_i}{P \times \max(\text{eff}_i)} \]

\[ eff_i = \frac{T_i}{T} \]

\[ IPC \]

\[ \# \text{instr} \]
Communication efficiency = $\mu$LB * Transfer

Serializations / dependences ($\mu$LB)

Measured using Dimemas
- An ideal network $\rightarrow$ Transfer (efficiency) = 1
Scaling model

Dimemas simulation with ideal target
- Latency = 0; BW = \infty

\[ \text{CommEff} = \muLB * \text{Transfer} \]

\[ \muLB = \frac{\text{max}(T_i)}{T_{\text{ideal}}} \]

\[ \text{Transfer} = \frac{T_{\text{ideal}}}{T} \]
Why scaling?

\[ \eta_\parallel = LB \times Ser \times Trf \]

CG-POP mpi2s1D - 180x120

Good scalability !! Should we be happy?

\[ \eta = \eta_\parallel \times \eta_{instr} \times \eta_{IPC} \]

Using Clustering to identify structure

Completed Instructions

IPC

Instructions Completed

Instructions Completed

Automatic Detection of Parallel Applications Computation Phases (IPDPS 2009)
Performance @ serial computation bursts

SPECFEM3D

WRF 128 cores

GROMACS

Asynchronous SPMD
Balanced #instr variability in IPC

SPMD
Repeated substructure
Coupled imbalance

MPMD structure
Different coupled imbalance trends
Projecting hardware counters based on clustering

Full per region HWC characterization from a single run

González J. et al, “Performance Data Extrapolation in Parallel Codes”. ICPADS '10
Integrating models and analytics

What if ....

... we increase the IPC of Cluster1?

... we balance Clusters 1 & 2?

PEPC

13% gain

19% gain
OpenMX (strong scale from 64 to 512 tasks)
Folding
Folding: Detailed metrics evolution

- Performance of a sequential region = 2000 MIPS

  Is it good enough?

  Is it easy to improve?
Folding: Instantaneous CPI stack

- Trivial fix. (loop interchange)
- Easy to locate?
- Next step?
- Availability of CPI stack models for production processors?
  - Provided by manufacturers?
“Blind” optimization

From folded samples of a few levels to timeline structure of “relevant” routines

Recommendation without access to source code

CG-POP multicore MN3 study

- Unbalanced MPI application
  - Same code
  - Different duration
  - Different performance
Methodology
Help generate hypotheses

Help validate hypotheses

Qualitatively

Quantitatively
First steps

Parallel efficiency – percentage of time invested on computation
  - Identify sources for “inefficiency”:
    • load balance
    • Communication /synchronization

Serial efficiency – how far from peak performance?
  - IPC, correlate with other counters

Scalability – code replication?
  - Total #instructions

Behavioral structure? Variability?

Paraver Tutorial:
Introduction to Paraver and Dimemas methodology
BSC Tools web site

www.bsc.es/paraver

- downloads
  - Sources / Binaries
  - Linux / windows / MAC

- documentation
  - Training guides
  - Tutorial slides

Getting started

- Start wxparaver
- Help → tutorials and follow instructions
- Follow training guides
  - Paraver introduction (MPI): Navigation and basic understanding of Paraver operation