High-level Data Analysis and Visualization

Xu Liu (Lead)
John Mellor-Crummey (Scribe)

Motivation

- Avoid duplication of development and optimization efforts from tool developers
- Develop best of breed
  - Analysis techniques
  - Data presentation/visualization methods
    - Calling context tree
    - Flame graphs for both profiles and traces
- Needs
  - Aggregation of performance data
  - Differential analysis
- Styles
- Code-centric
- Resource-centric
- Data-centric
- Time-centric

Challenges

- Multidimensional data is difficult to digest
  - Strategies
  - principal component analysis
  - Elide some data
- Optimization knowledge is difficult to acquire, especially source code modification
- Cannot always depend on compilers

Ideas

- Tag entities with semantic meta-data
- Tracing supplies important information unavailable otherwise
- Plug in model for Easy View?
  - Good way to provide a lot of data formats
    - For large-scale data, important to provide random-access to parts
  - Some data formats
  - Support for querying from disk
- Performance co-pilot that provides suggestions to a developer while writing code
- Need standard format for feedback or guidance
- Need standard format for metadata that describes the data
- If tool data is not ingested for display and analysis - can there be a mapping defined that can be structured for a new tool to provide data description, recipes for analysis/display and pointer to data
- Need metadata associated with metrics so we know what they mean

Strategies

- Top-down, hierarchical presentation and summary of metrics, e.g., decision tree
- Total resource consumption
  - Waste
  - Useful
  - Pattern recognition to identify bottlenecks (APART project)
    - https://www.researchgate.net/publication/2573802_Knowledge_Specification_for_Automatic_Performance_Analysis/link/0deec527764fcab435000000/download
- Pattern matching for optimization suggestions (GPA)
  - https://arxiv.org/abs/2009.04061
- W^3 model in Paradyn: paradyn.dvi (umd.edu)
- Providing guidance
  - Find bottlenecks
  - Describe bottlenecks
  - Summarizing behavior into known buckets, vs. other
- Dataset of code optimization prepared for code auto-optimization AI models
  - Autotuning may provide some data points
  - Data collection is challenging: hundreds of thousands of programs should be in the dataset
  - The dataset can be used for training AI models, as well as validate various performance tools