Capacity Metric for Chip Heterogeneous Multiprocessors

Mof Otoom
Yarmouk University, Jordan
June 26th, 2013
In this work we define and develop a new metric, Capacity.
Problem

- Existing metrics fail to properly evaluate and therefore rank the performance of emerging computer systems
  - The key difference between these systems and previous generations of computers is that they process multiple, variable heterogeneous workloads on heterogeneous multiprocessors in the service of single users

- This problem was discovered while investigating the idea of Workload Specific Processors (WSPs)
  - Designing single-user, multicore computers for patterns of workloads
    - In order to do this, there was a need to rank optimal performing designs for specific usage patterns
Chip Heterogeneous Multiprocessors

- We are targeting CHMs integrated into mobile devices such as iPhones.

At different times, different sets of apps arrive:
- Receiving a pic.
- Surfing MLB.com
- Opening gmail
- Listening to MP3

Performance is a vector of latencies/rates and is situational.

How does the presentation of jobs affect the response of the system?

For a variety of loading situations, which is better, Design A or Design B?
Primary Contributions

- Definition and development of the Capacity metric as a surface with dimensionality related to the number of demands processed by the CHM

- Techniques for analysis of the Capacity metric
Secondary Contributions

Definition and development of three foundations:
- a CHM model
- a multimedia cell phone example
- a Workload Specific Processor (WSP)

Definition of Workload Modes
- Definition and comparison of two approaches to workload mode identification at run time; The Workload Classification Model (WCM) and another model that is based on Hidden Markov Models (HMMs)

Development of a Demand Characterization Method (DCM) as a foundation for analysis of the Capacity metric
Outline

- Automobile Analogy
- Capacity Definition
- Contrast to Pareto Optimization
- Significance of Capacity Curve Shape
- Capacity Curve Shape Analysis
Outline

- Automobile Analogy
- Capacity Definition
- Contrast to Pareto Optimization
- Significance of Capacity Curve Shape
- Capacity Curve Shape Analysis
Automobile Analogy

- Capacity is to CHMs is what throughput is to pipelines
  - Latency is analogous to a craftsman
  - Throughput is analogous to an assembly line
  - Capacity is analogous to plants that contain multiple assembly lines
Automobile Example

The processing of multiple, variable heterogeneous workloads is analogous to the production of multiple types of automobiles in a production plant composed of multiple manufacturing pipelines.

Plant is not exactly matched all demands all the time

Which plant is better over a variety of demands?
Throughput

Throughput does not model the actual or useful work and is a function of supply only.

Throughput does not model the impact of the type of demand on the capabilities of multiple processor cores that are heterogeneous.

Throughput assumes output is a single stream. This results in a single unit value, described by its maximum.

assumes demand is constant
Automobile Example

Five different plant configurations – differ only in terms of line specialization

We assume a fixed plant resources

<table>
<thead>
<tr>
<th>Config.</th>
<th>Model-T Pipeline</th>
<th>Model-A Pipeline</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model-T</td>
<td>Model-A</td>
<td>Model-T</td>
</tr>
<tr>
<td>C1</td>
<td>6</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>C2</td>
<td>4</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>C3</td>
<td>10</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>C4</td>
<td>6</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>C5</td>
<td>9</td>
<td>2</td>
<td>9</td>
</tr>
</tbody>
</table>

C4 and C5 are homogeneous configurations
Average number of produced cars is different
Response curves cross
Different demand modes
Single-Valued Metrics

Demands are evaluated independently and measured during a short time unit. Demand always non-zero.

Throughput is normalized to the time interval.

Different averages result in different ranking, as well.

Response time of maximum production.

<table>
<thead>
<tr>
<th>Config.</th>
<th>Demand</th>
<th>Arithmetic</th>
<th>Harmonic</th>
<th>Geometric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AM</td>
<td>Rank</td>
<td>HM</td>
<td>GM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>17.8</td>
<td>3</td>
<td>18.8</td>
<td>18.8</td>
</tr>
<tr>
<td></td>
<td>20.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0,21)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>10.0</td>
<td>5</td>
<td>15.3</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td>32.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(16,0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>26.7</td>
<td>2</td>
<td>15.6</td>
<td>17.1</td>
</tr>
<tr>
<td></td>
<td>11.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>23.3</td>
<td>4</td>
<td>20.5</td>
<td>20.7</td>
</tr>
<tr>
<td></td>
<td>18.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>20.0</td>
<td>1</td>
<td>17.1</td>
<td>17.3</td>
</tr>
<tr>
<td></td>
<td>15.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ P(\tau) = (\tau / t_u) \times \text{avg} \]
Problem: Rank Order

- Different production sets result in different ranking
- Different averages result in different ranking, as well

<table>
<thead>
<tr>
<th>Config.</th>
<th>Demands</th>
<th>Arithmetic</th>
<th>Harmonic</th>
<th>Geometric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D1</td>
<td>D2</td>
<td>D3</td>
<td>D4</td>
</tr>
<tr>
<td>C1</td>
<td>(0.8)</td>
<td>(0.21)</td>
<td>(3.10)</td>
<td>(6.6)</td>
</tr>
<tr>
<td>C2</td>
<td>3.8</td>
<td>10.0</td>
<td>7.7</td>
<td>14.6</td>
</tr>
<tr>
<td>C3</td>
<td>10.0</td>
<td>26.7</td>
<td>15.0</td>
<td>10.0</td>
</tr>
<tr>
<td>C4</td>
<td>10.0</td>
<td>23.3</td>
<td>13.3</td>
<td>10.0</td>
</tr>
<tr>
<td>C5</td>
<td>7.8</td>
<td>20.0</td>
<td>10.0</td>
<td>6.7</td>
</tr>
</tbody>
</table>

Response time

Five different plant configurations
Problem: Second Order Effects

- Scheduling products to the best available line vs random scheduling

Throughput is equal, measured using a common unit, automobiles

Throughput cannot answer: Which design can achieve a specific demand (3T, 6A)?
Workload Specific Processors

- We are motivated by the concept of Workload Specific Processors (WSPs)
  - Designing CHMs for a variety of single-user usage patterns

- In order to do this, there was a need to rank optimal performing designs for specific usage patterns
Single-User Usage Patterns

- Multicore, mobile devices are being widely used by single individuals for comprehensive wireless access to the Internet, often processing simultaneous applications.

The user places different requirements on the system.
Existing Computing Categories

- **General Purpose Processors (GPPs)**
  - Inputs are sequential (sequenced by the processor)
    - No overlap
  - This leads to averaging performance

- **Application Specific Processors (ASPs)**
  - Input is a set of applications that persist over time and arrive periodically
  - designed given upfront specifications
  - Performance is to meet deadlines
    - Increasing computation power is wasteful
New Computing Era – Workload Specific Processors

- WSPs are to CHMs is what an ISA is to an RTL
  - Bridge CHMs to the way the system is used and will be programmed

- In contrast with GPP and ASP models, the input to WSPs is a timed set of heterogeneous applications, or workload modes
Workload Mode

- is a unique set of concurrently executing applications that includes external timing information
- opens up the possibility for tuning the system differently
- models the effects of task interactions on the performance of a CHM

The external events result from user–computer interaction and data arrival.

Performance is a vector and is situational.

The internal events arise from changes in data.
Workload Models

- Workload modes need to be predicted in advance, so that optimal optimizations can be selected at design time.
  - We present a method of workload mode identification that exploits some design-time knowledge and contrast it with an HMM.

- To evaluate our models we needed to build: a CHM model, a multimedia cell phone example, and a WSP model.

- Traditional metrics failed to identify optimal performing designs for specific usage patterns.
Related Work – Metrics

<table>
<thead>
<tr>
<th>Work</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matters</td>
<td>Matters</td>
</tr>
<tr>
<td></td>
<td>Power, Power per Area, and Utilization, etc</td>
</tr>
<tr>
<td>Doesn’t Matter</td>
<td>Latency, and Response Time, etc</td>
</tr>
</tbody>
</table>

None has considered breaking out the heterogeneity of workloads or has modeled heterogeneity impact on the performance of a CHM that processes a heterogeneous multi-channel workload.
Related Work – Workload Models

- Latency can be used to evaluate the performance of single untimed inputs
- Throughput can be used to evaluate the performance of multiple untimed homogeneous inputs
- Multiple heterogeneous timed inputs, or Workload Modes, need a structurally different metric

<table>
<thead>
<tr>
<th>Concurrency / Type</th>
<th>Timing</th>
<th>Un-timed</th>
<th>Timed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td></td>
<td>Benchmarks, Traces</td>
<td>Stochastic, Statistics</td>
</tr>
<tr>
<td>Multiple</td>
<td>Homogeneous</td>
<td>Benchmarks, Behavioral Graphs</td>
<td>Stochastic</td>
</tr>
<tr>
<td></td>
<td>Heterogeneous</td>
<td>Benchmarks, Behavioral Graphs</td>
<td><strong>Workload Modes</strong></td>
</tr>
</tbody>
</table>
Outline

- Automobile Analogy
- Capacity Definition
- Contrast to Pareto Optimization
- Significance of Capacity Curve Shape
- Capacity Curve Shape Analysis
Thinking about automobile production is what led to the definition of Capacity.
Capacity

- Capacity is a surface that shows the feasibility of production over a heterogeneous set of types for a given plant over some interval of time.

- Capacity curves initially are generated using measurement:
  - Includes the overhead of sharing resources between different demands.

- Dominate by points A or C, whichever the model of interest.

- Unfeasible, efficient, feasible but inefficient.

- Shape has a meaning.

- Diagram:
  - Model-A vs Model-T
  - Points A, B, C, D, E, F
  - Area of interest
  - Deadlines
  - Feasible, unfeasible, efficient, inefficient.
Outline

- Automobile Analogy
- Capacity Definition
- Contrast to Pareto Optimization
- Significance of Capacity Curve Shape
- Capacity Curve Shape Analysis
Pareto Optimization

- A change in the design that makes at least one output model better off without making any other outputs worse off is called a Pareto improvement
  - A combination is defined as Pareto efficient or Pareto optimal when no further Pareto improvements can be made

- The most efficient production for a model is when the production of other model(s) is zero
  - Similar to the definition of throughput
  - The most efficient production of all models are then shown on a curve
What is the impact of increasing/decreasing the production of one model on the performance of a design?

What are the effects of resource sharing on the production of a design?
Outline

- Automobile Analogy
- Capacity Definition
- Contrast to Pareto Optimization
- Significance of Capacity Curve Shapes
- Capacity Curve Shape Analysis
Significance

We found that the shape of Capacity curves can be more significant than magnitude in evaluating performance.
Idealized Capacity Forms

- **1core_dynamic_zero_overhead**
  - A single core system that ideally processes unlimited combinations of text and JPEGs

- **2core_static_decoupled**
  - A homogeneous two-core processor system in which tasks are mapped to restricted cores

- **2core_static_coupled**
  - Similar as 2sd but includes the cost of resource sharing

No quantization effects and zero switching penalty

Completely, ideally decoupled

Double the capacity of a uP
**Idealized Capacity Forms**

- **2core_dynamic_decoupled**
  - A homogeneous two-core processor system in which tasks are dynamically mapped to **either** of two processor resources.

- **2core_dynamic_coupled**
  - Similar as 2dz but includes the cost of resource sharing.

- **We also expect to find multi-modal shapes**

  - No quantization effects and zero switching penalty.
  - Completely, ideally decoupled.
  - Double the capacity of a uP.
Experimental Setup

Multiple tasks concurrently execute on multiple PEs

- JPEG Images
- Text Files
- Flash Frames

Different complexities and working set sizes

L-2 cache size

Philips PNX1700

The Modeling Environment for Software and Hardware simulator**

32-bit, 200-Mhz bus

Static and dynamic schedulers

Burst width, communication bandwidth

ARM926EJ-S

Blackfin533

Capacity

Processing, Communication, and Overhead Modeling

- The performance of each task is done through application profiling

- We model the overhead due to:
  1. gathering global chip state, (2) evaluating and sending scheduling decisions, (3) bus contention, (4) cache memory misses, and (5) task migration

- The Modeling Environment for Software and Hardware (MESH) is also capable of evaluating communication and memory contention
Results – Burst vs Communication

The B2 curve is the more ideal shape for the overall design goals.

Overall, burst width favors more unbalanced loads – it facilitates dynamic scheduling.

Both approaches the straight line of a perfect scheduler, even though the processing of input channels is coupled.

Overall, increasing communication bandwidth decouples system resources.
Results – Processor Type

DSP processors perform better than GPP and Media when a mix of text and JPEG is being produced.

Almost equal

Performance cannot overcome overhead

Almost equal

More GPPs

DSPs are better for text

Media is better for JPEGs

Arch4: 2G, 4D, 3M
Arch3: 2G, 2D, 6M
Arch2: 2G, 2D, 6M
Arch3: 1G, 4D, 6M

The two curves differ not just in terms of magnitude but also shapes.

Overall, increases the number of processors of specific type shifts curve towards products that take more advantage of this type.
Results – Time Interval

- Throughput is presumed to be invariant over arbitrary time intervals.

- By contrast, capacity surfaces can change with interval size.
  - Overhead due to tasks interactions.

Doubling the time window does not result in double capacity.

Increasing time favors homogeneous inputs.
Outline

- Automobile Analogy
- Capacity Definition
- Contrast to Pareto Optimization
- Significance of Capacity Curve Shapes
- Capacity Curve Shape Analysis
The best design using single-valued metrics:

- **Average:**
  - AM: Design B
  - HM: Design A
  - GM: Design B
- **Difference Area is zero**
  - All Designs are equal
- **Maximums:** Design B

Design A: 1GPP, 4DSPs
Design B: 2GPPs, 4Media
Design C: 6DSPs
Our Approach

- We developed a Demand Characterization Method that characterizes the demand of a specific usage pattern in the form of a curve.

- We overlay demand curves over Capacity curves of different architectures to compare their performance and identify optimal designs.

Different types of demand curves

Design A does not meet these demands

Capacity curve of Design A
Example

Design B does not meet all demands – it does not show excessive capacity as well.

Design C has met all occurrence demands but not all maximum.

Design C seems the best, by contrast to other metrics.
The optimal design rank may change for different demand types.

Again, Design C is the best performing design, by contrast to single-valued metrics.

Design B is the second best design.
Another Scenario

- The optimal design rank may change for different usage patterns
  - This scenario favors unbalanced loads

Design B is the best performing design for this scenario.
Overhead Modeling

- Our Capacity models the overhead of global resource interactions.
- The effects of overhead is what makes Capacity curves multimodal.

<table>
<thead>
<tr>
<th>Demand (JPEGs, GIFs)</th>
<th>Overhead (Mega Cycles)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chip Level Scheduling</td>
</tr>
<tr>
<td>(0,10)</td>
<td>0.342</td>
</tr>
<tr>
<td>(3,8)</td>
<td>0.512</td>
</tr>
<tr>
<td>(6,9)</td>
<td>0.444</td>
</tr>
<tr>
<td>(9,7)</td>
<td>0.433</td>
</tr>
<tr>
<td>(12,5)</td>
<td>0.489</td>
</tr>
<tr>
<td>(15,4)</td>
<td>0.549</td>
</tr>
<tr>
<td>(18,6)</td>
<td>0.512</td>
</tr>
<tr>
<td>(21,2)</td>
<td>0.501</td>
</tr>
<tr>
<td>(24,0)</td>
<td>0.449</td>
</tr>
</tbody>
</table>

Overhead of homogeneous inputs is lesser.

Capacity does not overcome the increase in overhead.
Conclusions

- We introduced a Capacity metric to evaluate the performance of a production system composed of multiple, heterogeneous production units and used to simultaneously produce multiple types.

- We have shown using analytical techniques how shapes of the Capacity curves must be used to properly rank and classify the performance of systems that process multi-channel workloads.
Conclusions

- We developed three foundations that we used to develop our Capacity metric: a CHM model, a multimedia cell phone example, and a WSP.

- We defined workload modes, derived from models of single-user access patterns, as a means of design orientation and performance optimization for future single-user CHMs.
Future Work Directions

- Multi-Dimensionality Analysis
- User Profiling
Future systems are expected to have many channels of inputs
- Graphical comparisons of the capacity metric will not scale beyond three dimensions

Surface analysis will be done in order to extract features in a feature vector
- Initially includes: [Magnitude, Convex/concavity, Modes, Slope]
Since Capacity metric was originally emerged from the need to optimize CHMs to workload modes that result from individual users using multicore mobile devices, single-user usage profiling is essential for discovering real usage patterns or workloads. In this way, Capacity curves can be used to identify for each user where performance does not meet, exactly matches, or excessively meets his needs and by how much.
Thank You!

Questions
Additional Results – Multiple Effects

For the same system when communication bandwidth increased:

- We increased working sets size.
- We increased cache size by 1.5 times.
- Curve endpoints are equal.
- System changes from favoring unbalanced loads to favoring balanced loads.
The input now is GIFs and JPEGs.

System changes from favoring balanced loads to an ideal system that switch the production between tasks with approximately zero cost.

Increasing time results in almost double capacity for all combinations because JPEGs and GIFs have same memory and processing patterns.