Virginia Tech
Invent the Future

An iso-energy-efficient approach to scalable system power-performance optimization.

Leon Song, Matthew Grove, Kirk Cameron
SCAPE Lab, Virginia Tech

August 24, 2011

Virginia Tech
Invent the Future

Background

- Since 1992 performance has increased 10,000 fold while performance per watt only improved 300 fold.
- Energy efficiency is now key to HPC system design.
- In order to continue to scale we must address the energy problem.

VirginiaTech
Invent the Future

SCAPE Lab

- Focus on power and performance.
- Co-founded the Green500.
- We dismantle your expensive HPC nodes and directly instrument hardware (hopefully without releasing the magic smoke).

Talk Focus

- Motivation for producing a new model.
- Gathering the model input parameters.
- What can you do with the model.
- General things we have learned from using the model.

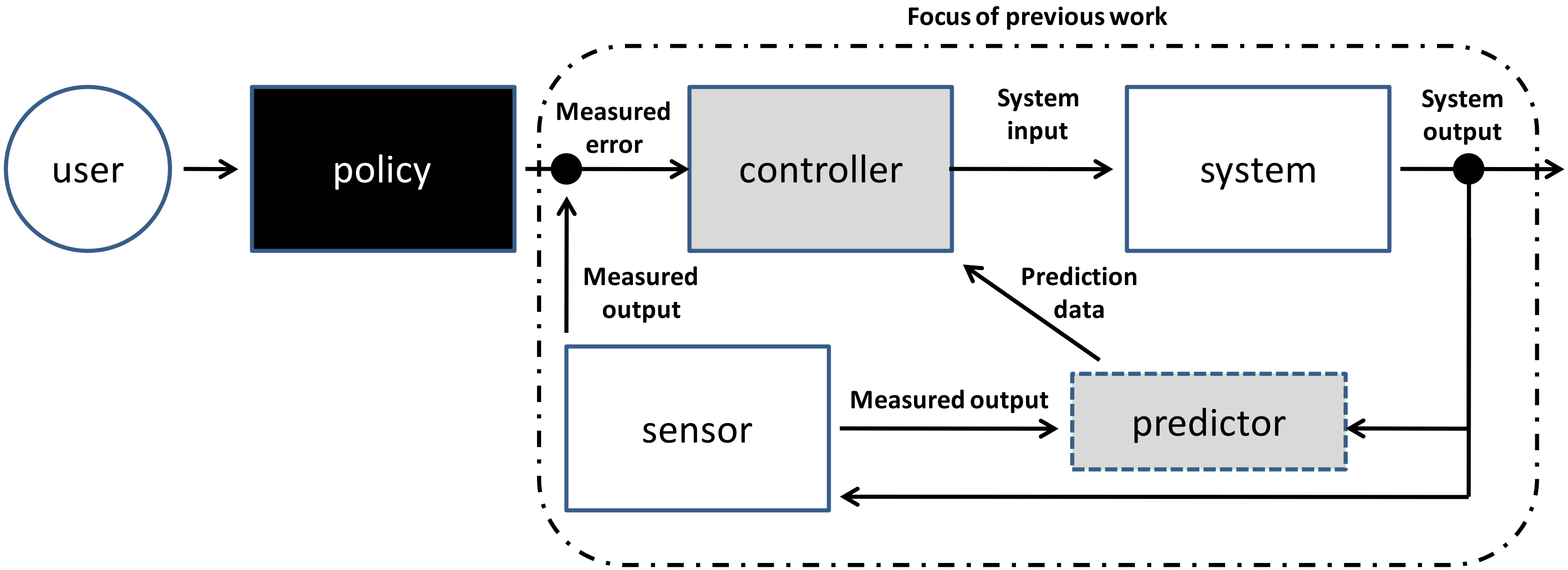
Problem

- We do not fully understand the impact of system–level power management on application performance.
- What is the root cause of any performance or power changes?

VirginiaTech Invent the Future

Current Approaches

- The majority of the work focuses on power mode **predictor** and **controller** design.



VirginiaTech
Invent the Future

Modeling vs Observing

- We want to be able to predict ahead of time what will happen if we alter anything about how a job is run.
- Such as changing the resources allocated to the job or altering the power management strategy.

Use Cases

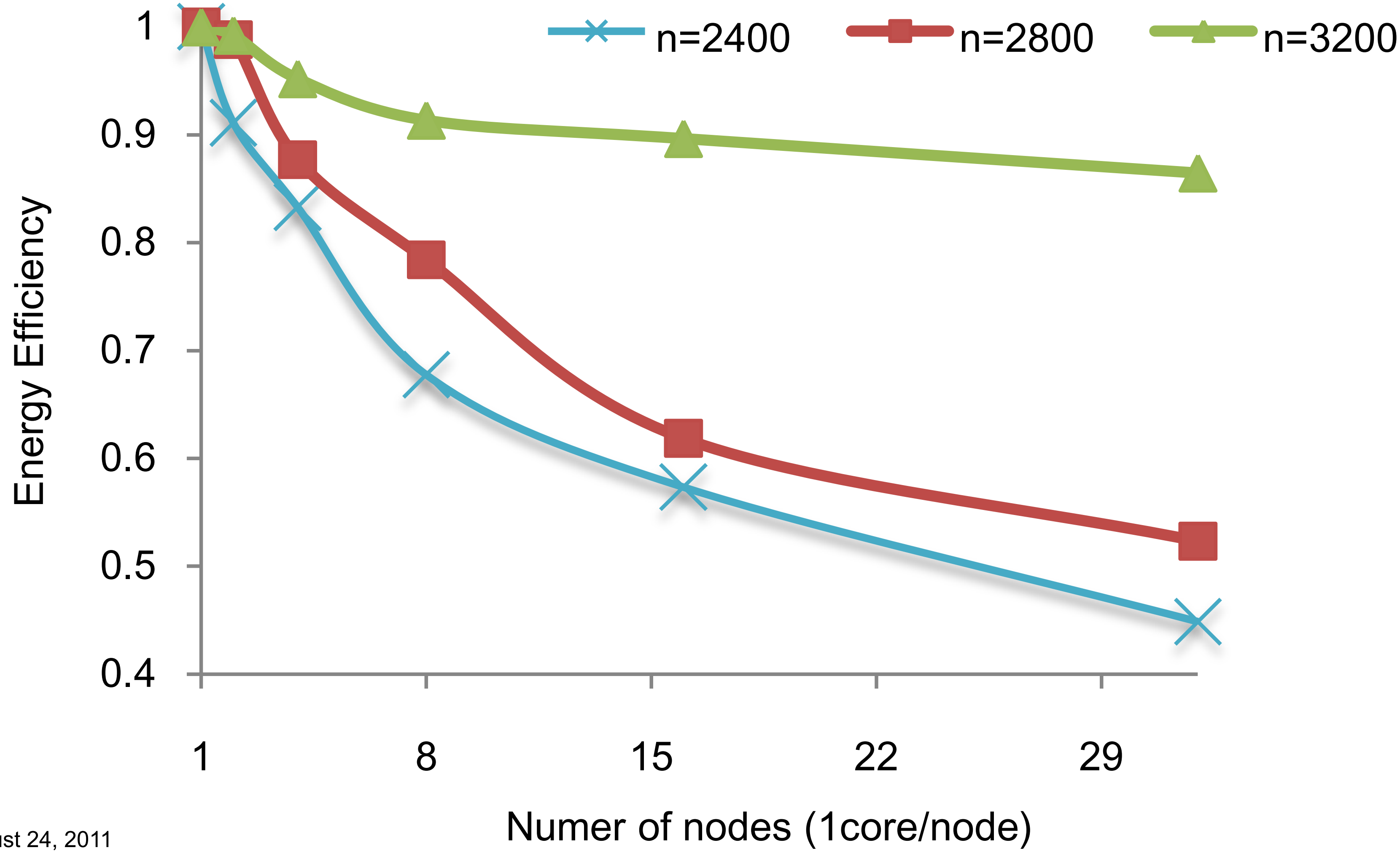
- Enable users to explain an observed efficiency.
- Determine the root cause of the inefficiency.
- Help a system designer identify inefficiencies in system or algorithm design.

System Energy Efficiency

- We can illustrate the effect of scaling problem size on system efficiency with a simple experiment.
- We apply Cannon's algorithm to varying problem sizes with the CPU in a fixed power mode (frequency) whilst varying the system size.

VirginiaTech
Invent the Future

Problem Size



Scaling Problem Size

- The graph shows that for this simple example scaling system size with problem size can increase efficiency.

Approach

- Build an analytical model for both power and performance to gain insight into how they interact.
- The goals for the model:
  - Practical (usable)
  - Accurate
  - Useful

Iso-Energy-Efficiency (I-I-E)

- Quantitatively model the interactive effects of power and performance on clusters.
- Addresses two key points:
  - Predict total energy consumption.
  - Model how energy efficiency is affected by changing parameters such as CPU frequency.

Methodology

- Run the application and gather input parameters.
- Build the Energy model, combining:
  - Performance and Power models.
- Find optimal values for system energy efficiency.

I-I-E Parameters

- There are 29 inputs in the model, loosely grouped:
  - **Machine Dependent,** e.g. number of nodes.
  - **Time Related**, e.g. average time to send a message.
  - **Power Related**, e.g. average CPU power in idle state.

Case Studies

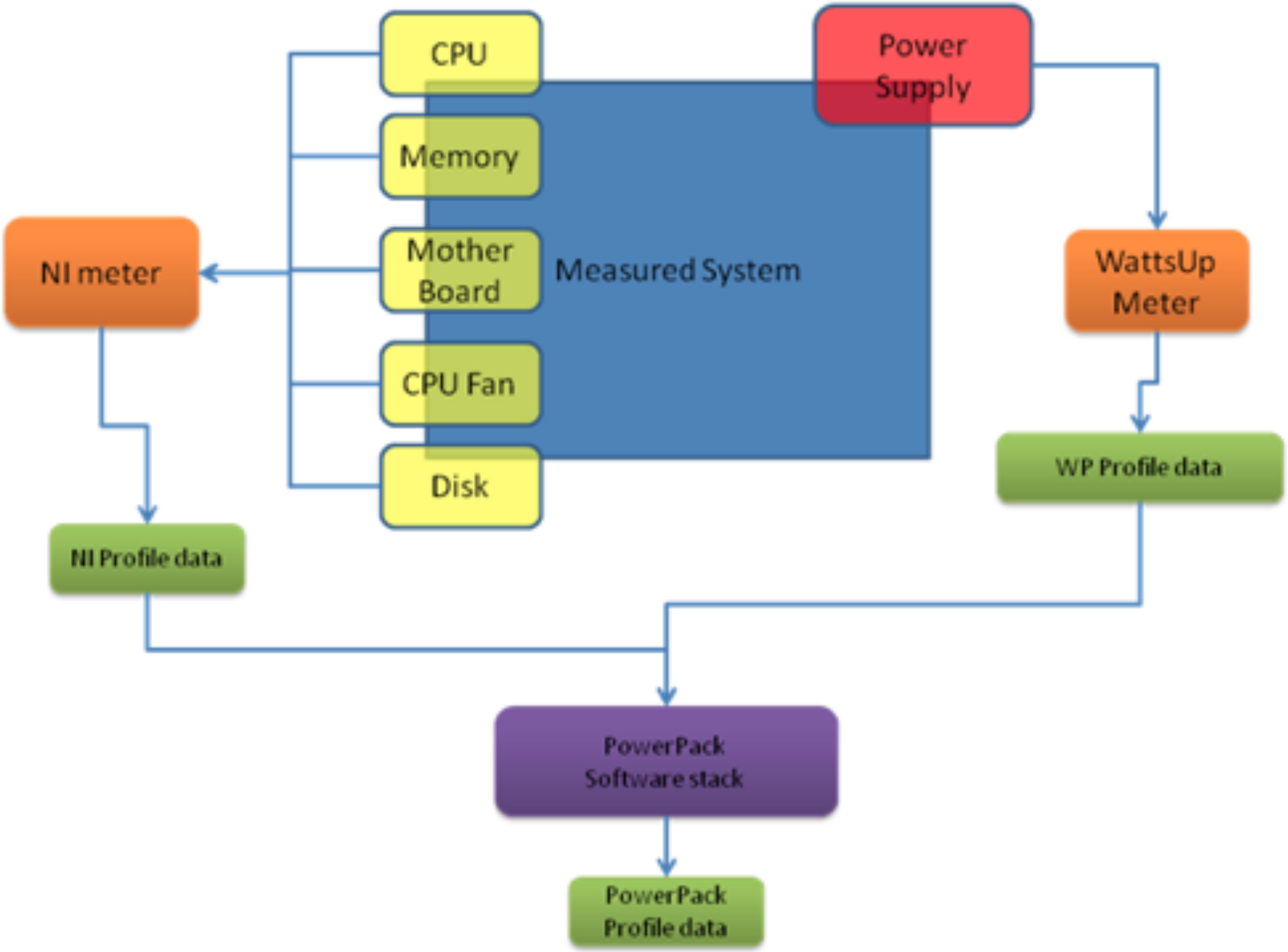
- Our instrumented power aware clusters were used.

Cluster	System size	Processor	Memory	L1 cache	L2 cache	Interconnection	frequency
SystemG	325 Mac Pro nodes	two quad-core 2.8 GHz Intel Xeon processor	8GB RAM	32KB	Shared, 6MB	Mellanox 40Gbytes/sec InfiniBand	2.8 and 2.4 GHz
Dori	8 blades	AMD Opteron dual core dual processor	6GB RAM	64KB	Shared, 1MB	1Gbytes/sec Ethernet	1.8, 1.6, 1.4, 1.2, 1.0 GHz

Collecting Parameters

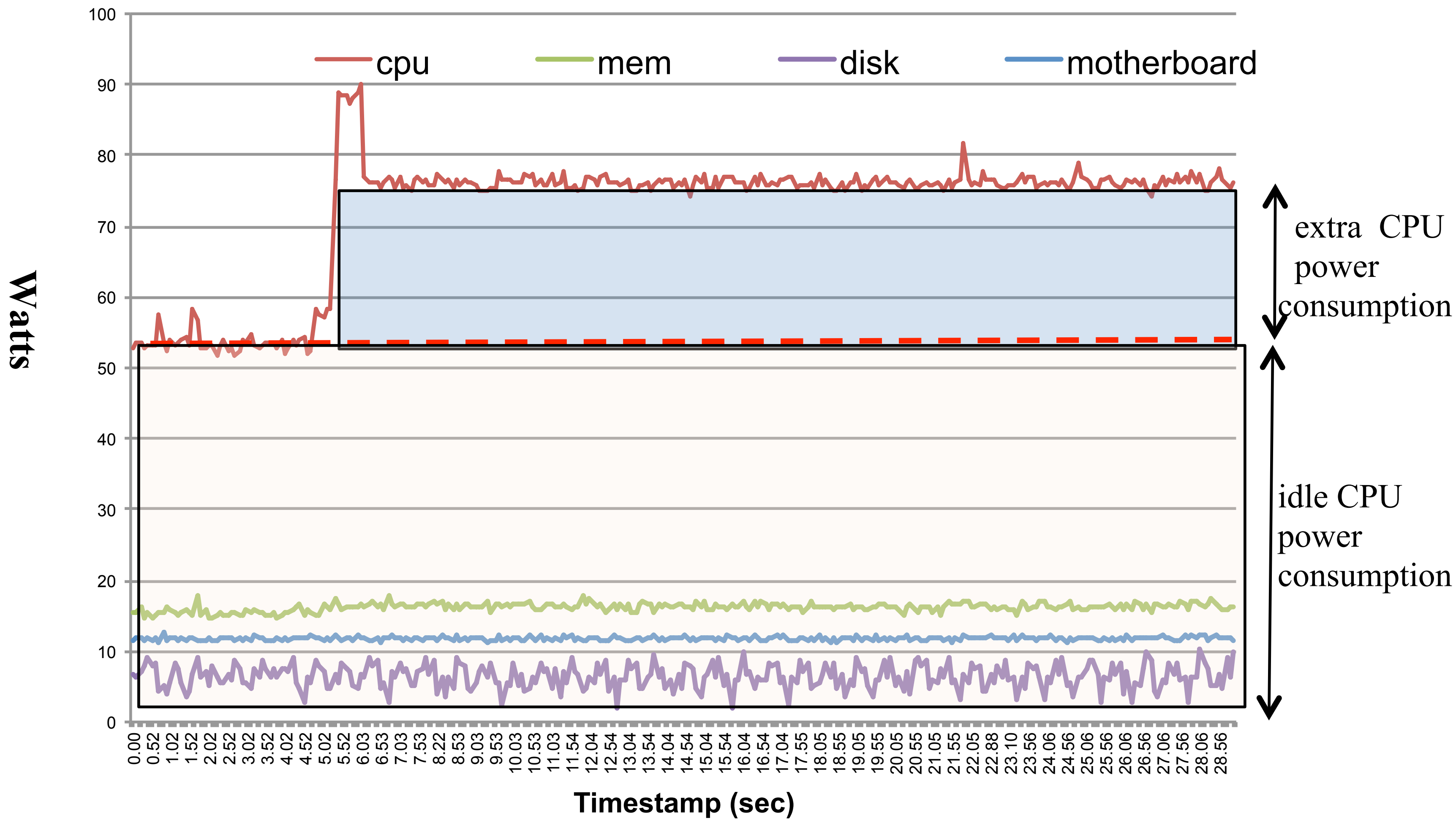
- **Perfmon+libpfm4.0**: Hardware counters
- **PowerPack 3.0**: Power
- **MPPTest**: MPI
- **LMbench**: Memory
- **/proc/stat**: IO

PowerPack



August 24, 2011

PowerPack Data



August 24, 2011

PowerScale

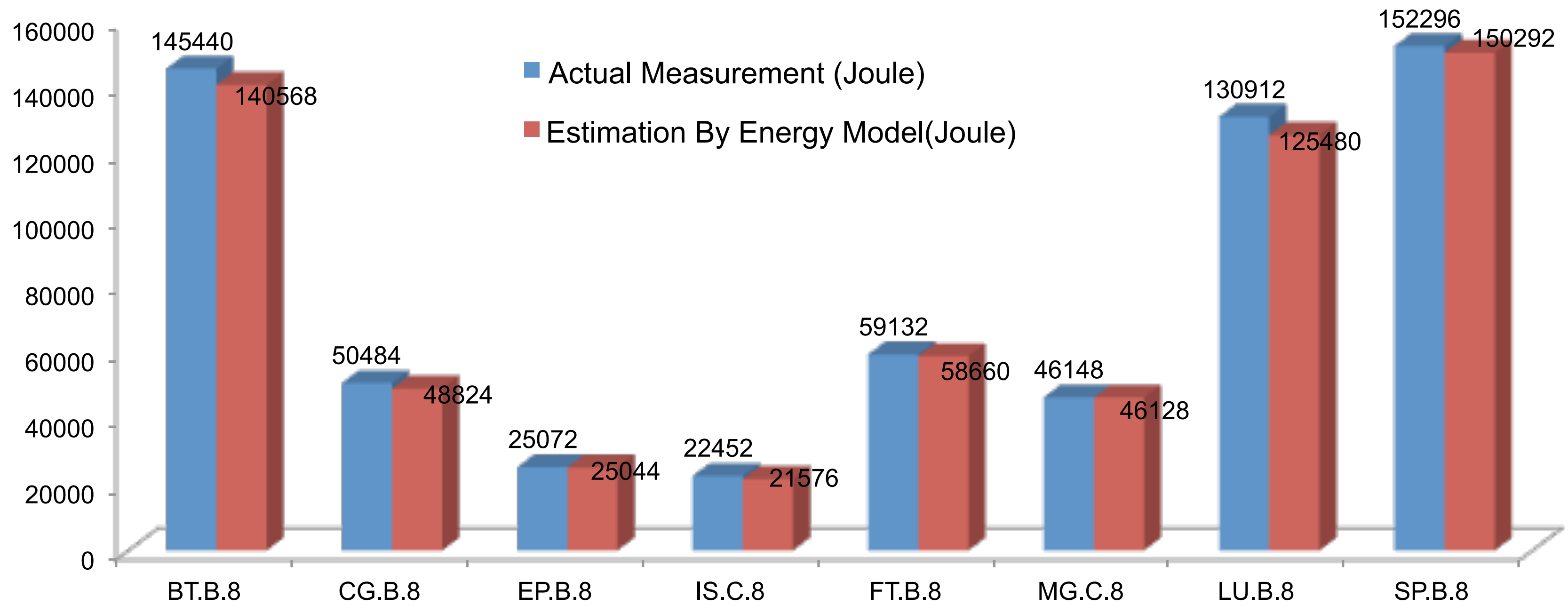
- Manually gathering the parameters was very labor intensive and error prone.
- We developed a runtime called PowerScale to automate this part of the work.

Measuring Accuracy

- We ran the NAS parallel benchmark suite on Dori and SystemG.
- We compared the energy consumption as predicted by the model to actual consumption as measured by PowerPack.

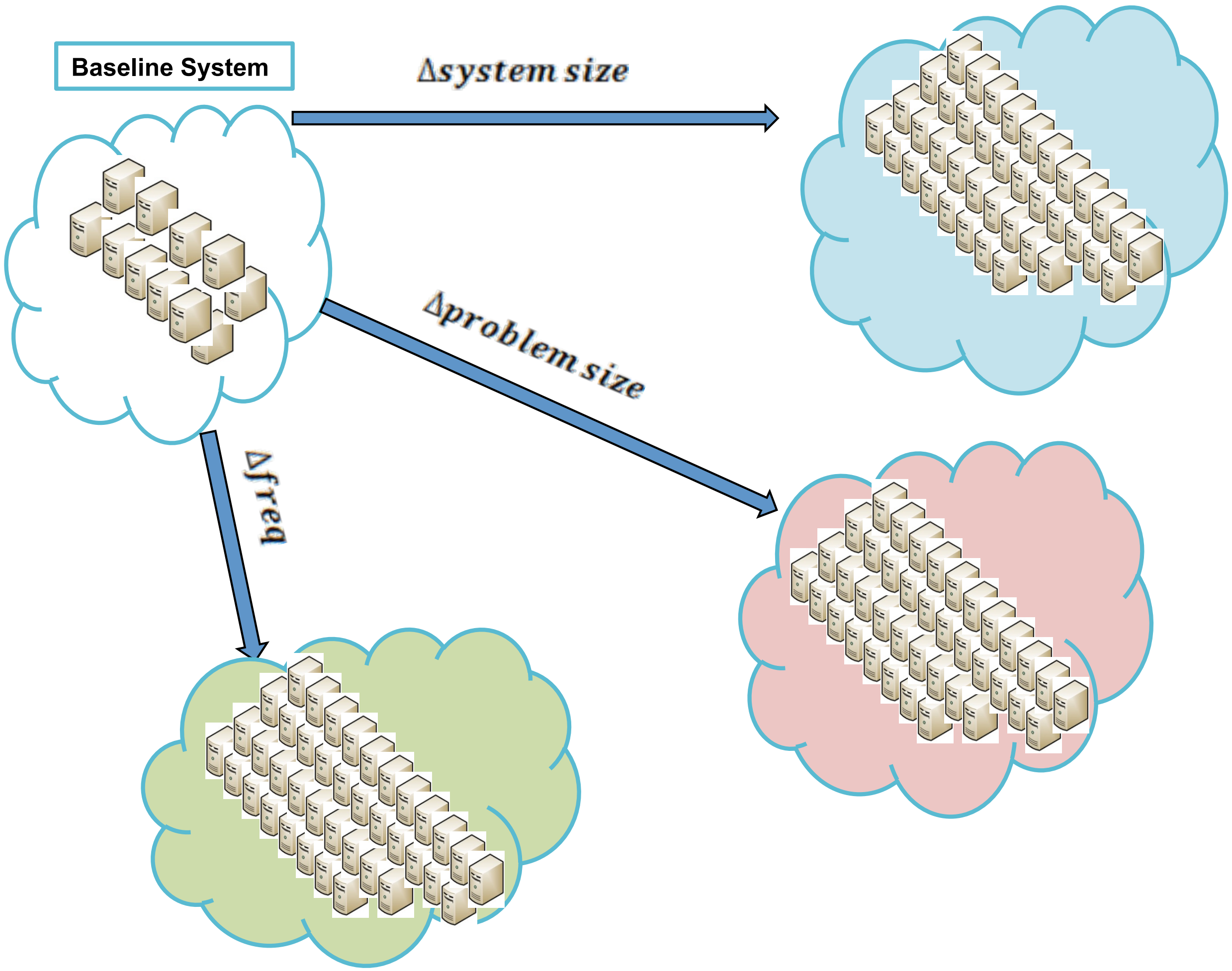
Dori NAS Accuracy

- Model accuracy >95% in all benchmarks. 8 nodes fixed frequency.



August 24, 2011

I-E-E Uses



VirginiaTech
Invent the Future

Applying the Model

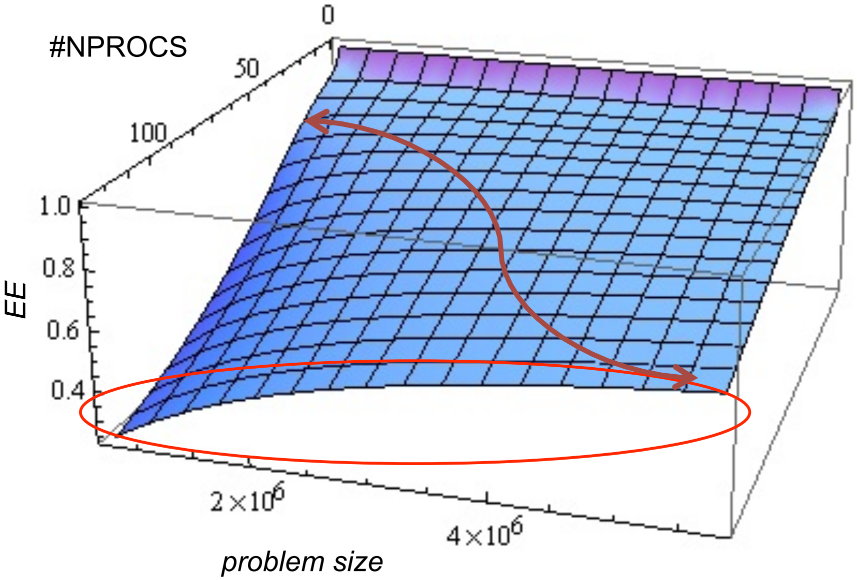
- Iso-Energy-Efficiency is still very new (introduced in IPDPS 2011).
- We wanted to put it to practical use.
- Use the model to determine appropriate efficiency values for problem size and power scaling modes on clusters.

Case Studies

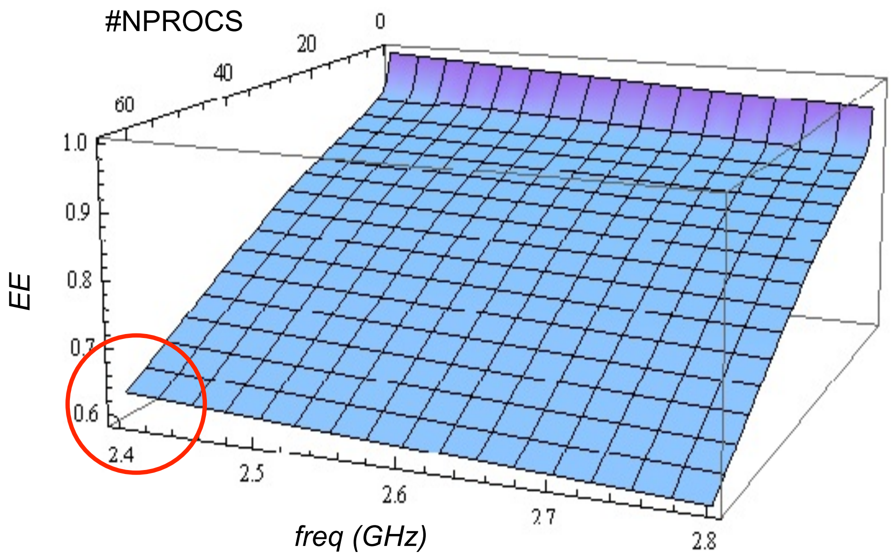
- We have analyzed several benchmarks (see papers).
- We will look at Fourier Transform (FT) and Conjugate Gradient (CG) from the NAS parallel benchmark.
- FT is communication intensive with dominating communication for some execution phases. CG is more computationally intensive.

Predicting for FT

FT's system-wide energy efficiency with p and n as variables



FT's system-wide energy efficiency with p and f as variables



August 24, 2011

FT Observations

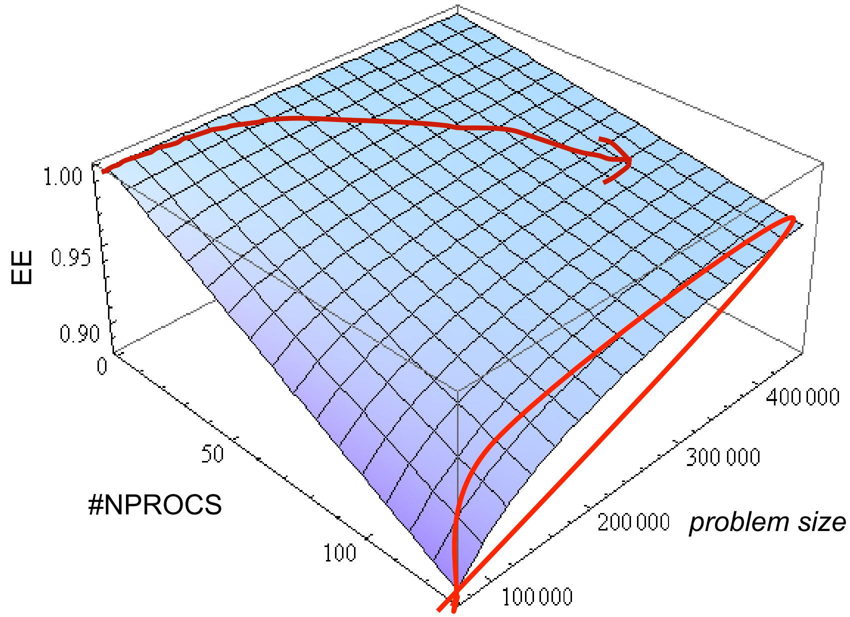
- Problem size scaling under fixed frequency is effective in maintaining overall system energy efficiency.
- CPU frequency scaling does not drastically effect the efficiency.
- Conclusion: Scale number of nodes and problem size simultaneously.

VirginiaTech
Invent the Future

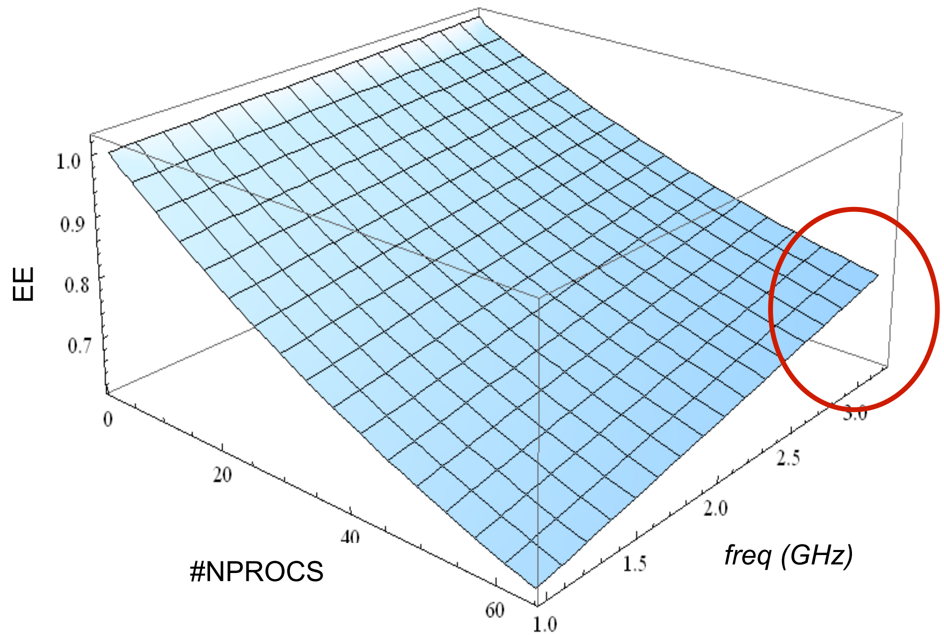
Predicting for CG

- More CPU intensive than FT.

CG's system-wide energy efficiency with p and n as variables



CG's system-wide energy efficiency with p and f as variables



CG Observations

- The energy efficiency declines as more parallelism is added.
- Energy efficiency can be maintained by scaling problem size.
- CPU frequency has more impact than with FT because of the lower communication to computation ratio.
- Conclusion: Scale problem size, nodes and CPU frequency.

VirginiaTech
Invent the Future

Conclusions

- **Practical** (usable), although it is made easier if you have a tool for automating measuring the parameters.
- **Accurate** within 5%.
- **Useful** for predicting total system energy consumption and allows ‘what if’ analysis.

Problem Size Scaling

- **Pros**:
  - Large range to scale gives flexibility.
- **Cons:**
  - Does not fit problems with limited input data or limited system resources.

Frequency Scaling

- **Pros**:
  - Potential to save a lot of energy.
- **Cons:**
  - Limited frequencies can restrict the rate of system energy improvement.
  - Does not improve system utilization.

Future Work

- Automate the analysis part of the model that happens after running PowerScale.
- Additionally make a simpler version of the model (sacrificing some accuracy) in order to make it easier to apply.

Virginia Tech
Invent the Future

GPU

- We are interested in extending the model to work with heterogeneous architectures such as the increasingly popular GPU.
- We do not currently instrument PCI cards as part of PowerPack. How can we get the energy consumption for a single GPU?

Questions

- Leon Song
- s562673@cs.vt.edu
- http://scape.cs.vt.edu/