The Challenge of Scale (Reprised)

Fault Tolerance, Scaling and Adaptability

Dan Reed
Dan_Reed@unc.edu
Renaissance Computing Institute
University of North Carolina at Chapel Hill

http://lacsi.rice.edu/review/slides_2006/

LACSI

Acknowledgments

- Staff
  - —Kevin Gamiel
  - —Mark Reed
  - —Brad Viviano
  - —Ying Zhang
- Graduate students
  - —Charng-da Lu
  - —Todd Gamblin
  - —Cory Quamman
  - —Shobana Ravi
- LANL and ASC insights
  - —a long, long list of people

renci
renaissance computing institute

LACSI

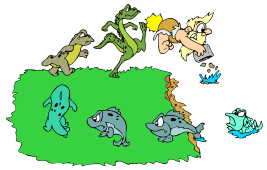
LACSI Impacts

- Market forces and laboratory needs
  - —multicore chips and massive parallelism
    - – capability and capacity systems
  - —power budgets ($) and thermal stress
    - – economics and reliability
- Tools and systems haven't kept pace
  - —scale, complexity, reliability and adaptation
- *Making large systems more usable (our focus)*
  - —*scale, measurement and reliability*
  - —*power management and cooling*
  - —*prediction and adaptation*
- Federal policy initiatives
  - —June 2005 PITAC computational science report (chair)
    - – "Computational Science: Ensuring America's Competitiveness"
  - —Computing Research Association (CRA) (chair, board of directors)
    - – Innovate America partnership

LACSI

LACSI Research Evolution

- At last year's review
  - —application fault resilience
  - —large-scale system failure modes
  - —HAPI health monitoring toolkit
  - —uniform population sampling
- This year
  - —AMPL stratified sampling toolkit
  - —Failure Indicator Toolkit (FIT)
  - —extended temperature/power measurements
  - —SvPablo application signature integration
  - —power-driven batch scheduling
- Research agenda driven by ASC challenges
  - —scale, performance and reliability



LACSI

You Know You Are A Big System Geek If …

- You think a $2M cluster
  - —is a nice, single user development platform
- You need binoculars
  - —to see the other end of your machine room
- You order storage systems
  - —and analysts issue "buy" orders for disk stocks
- You measure system network connectivity
  - —in hundreds of kilometers of cable/fiber
- You dream about cooling systems
  - —and wonder when fluorinert will make a comeback
- You telephone the local nuclear power plant
  - —before you boot your system

LACSI

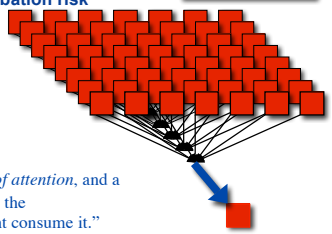
The Rise of Multicore Chips

- Intrachip parallelism
  - —dual core is here
    - – Power, Xeon, Opteron, UltraSPARC
  - —quad core is coming in just months …
    - – Intel, AMD, IBM, SUN
  - —Justin Ratter (Intel)
    - – "100's of cores on a chip in 2015"
- "Ferrari in a parking garage"
  - —high top end, but limited roadway
- Massive parallelism is finally here
  - —tens and hundreds of thousands of tasks

LACSI

Scalable Performance Monitoring

- Scalable performance monitoring
  - —summaries, space efficient but lacking temporal detail
  - —event traces, temporal detail but space demanding
- At petascale, even summaries are challenging
  - —exorbitant data volume (100K tasks)
  - —high extraction costs, with perturbation risk
- Tunable detail and data volume
  - —application signatures (tasks)
    - – selectable dynamics
  - —stratified sampling (system)
    - – adaptive node subset



"… *a wealth of information creates a poverty of attention*, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it."
Herbert Simon

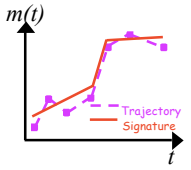
LACSI

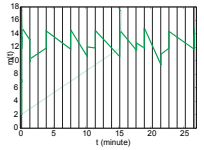
Compact Application Signatures

- Motivations
  - —compact dynamic representations
  - —multivariate behavioral descriptions
  - —adaptive volume/accuracy balance
- Polyline fitting
  - —based on least squares linear curve fitting
    - – measurement at user markers
  - —curves are computed in real-time
- Signature comparison
  - —degree of similarity (DoS) of q wrt p

$$\max\left(1-\frac{\int |p(t)-q(t)|\,dt}{\int p(t)dt},0\right)$$

- SvPablo integration
  - —marker selection inside GUI
  - —data capture library (DCL) signature generation
  - —signature browsing and comparison
- Adaptive measurement control





Source: Charng-da Lu (SC02 Best Student Paper Finalist)

LACSI

Sampling Theory: Exploiting Software

- SPMD models create behavioral equivalence classes
  - —domain and functional decomposition
- By construction, …
  - —most tasks perform similar functions
  - —most tasks have similar performance
- Sampling theory and measurement
  - —extract data from "representative" nodes
  - —compute metrics across representatives
  - —balance volume and statistical accuracy
- Estimate mean with confidence 1-α and error bound *d*
  - —select a random sample of size *n* from population of size *N*

$$n \geq N\left[1+N\left(\frac{d}{z_\alpha S}\right)^2\right]^{-1}$$

  - —approaches $\frac{z_\alpha S}{d}$ for large populations

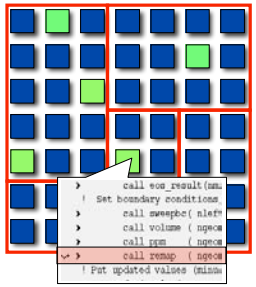
Sampling Must Be Unbiased!

Source: Todd Gamblin

LACSI

Adaptive Performance Data Sampling

- Simple case
  - —select subset *n* of N nodes
  - —collect data from the *n*
- Stratified sampling (multiple behaviors)
  - —identify low variance subpopulations
  - —sample subpopulations independently
  - —reduced overhead for same confidence
- Metrics vary over time
  - —samples must track changing variance
    - – number and frequency
  - —number of subpopulations also vary
- Sampling options
  - —fixed subpopulations (time series)
  - —random subpopulations (independence)
- Adaptive measurement control
  - —fix data volume (variable error)
  - —fix error (variable data volume)

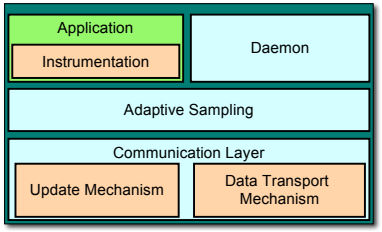


Source: Todd Gamblin

LACSI

AMPL Framework

- AMPL
  - —Adaptive Performance Monitoring and Profiling On Large Scale Systems
  - —SvPablo and TAU integration
  - —Multiple performance data sources (PAPI and others)



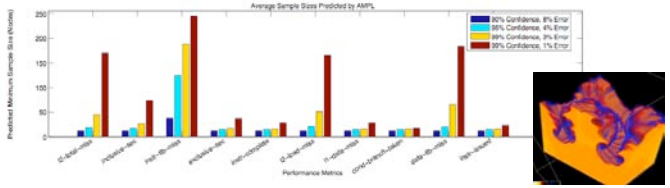

```
SampleWindow = 5.0
WindowsPerUpdate = 4
UpdateMechanism = Subset

Group {
     Name = "Adaptive"
     Members = 0-127
     Confidence = .90
     Error = .03
}
Group {
     Name = "Static"
     SampleSize = 30
     Members = 128-255
     PinnedNodes = 128-137
}
```

Source: Todd Gamblin

LACSI

sPPM Sampling Results



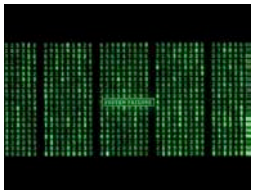
- PAPI counter sampling
  - —5-14% overhead at 90% confidence and 8% accuracy
  - —7-14% overhead at 99% confidence and 1% error
    - – low variance metrics

Source: Todd Gamblin

LACSI

Execution Models and Reliability

- There are many execution models
  - —parameter space exploration
  - —single program, multiple data (SPMD)
  - —master/worker and functional decomposition
  - —dynamic workflow
    - – data and condition dependent execution
- Each amenable to different reliability strategies
  - —need-based resource selection
  - —over-provisioning
    - – SETI@Home model
  - —checkpoint/restart
  - —algorithm-based fault tolerance
  - —library-mediated over-provisioning



LACSI

Machine Room Microclimate

- Sensors for machine rooms
  - —multiple locations
    - – air ducts, racks, servers, …
  - —multiple modes
    - – vibration, temperature and humidity
- Sensor options
  - —UC Berkeley/Crossbow motes
  - —WxGoos network sensors
- Infrastructure coupling
  - —HAPI for integrated data capture
  - —AMPL for statistical sampling
  - —FIT for failure model generation
  - —SvPablo for application instrumentation
- Rationale
  - —micro-environment analysis
  - —thermal gradients and equipment placement

Source: Shobana Ravi/Brad Vivano

LACSI

A Tale of Three Clusters

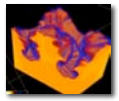
- Old, homemade (Dell)
  - —standard Dell towers
  - —1 GHz Pentium III dual processor nodes
  - —multiple rows of eight nodes
  - —GigE interconnect
- Clustermatic (Linux Labs)
  - —one 42U rack
  - —2 GHz Opteron dual processor nodes
  - —16 nodes plus head node
  - —Infiniband and GigE interconnects
- Vendor (Dell)
  - —17 standard racks, plus 4 network racks
  - —512 3.6 GHz Xeon dual processor nodes
  - —Infiniband interconnect

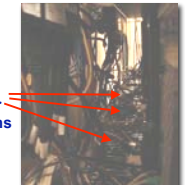
Source: Shobana Ravi

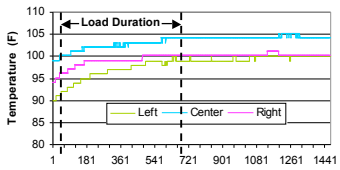
LACSI

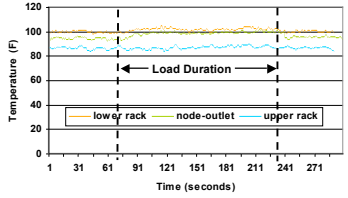
Loading and Monitoring Details

- UC Berkeley/Crossbow motes
  - —temperature measurements
- Measurement locations
  - —air outlet on each node
- Benchmark
  - —sPPM
- Observations
  - —rack cooling (or its lack) really matter





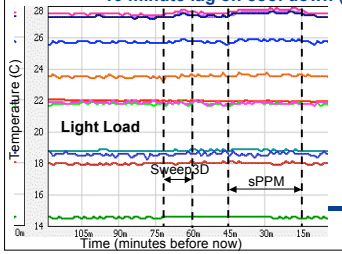




Source: Shobana Ravi

Clustermatic Temperature Profile

- WxGoos hardware
  - —temperature, power, humidity, …
- Measurement locations
  - —air outlets, sensors on rack door
- Multiple benchmarks
  - —sPPM and Sweep3D (multiple data sets)
  - —~10 minute lag on cool down (larger data)



WxGoos Sensors

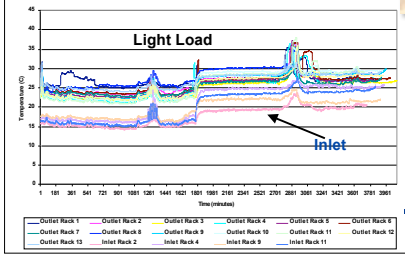
Source: Shobana Ravi

LACSI

Large Cluster: Top500 Benchmarking

- UC Berkeley/Crossbow motes
  - —temperature measurements
- Measurement locations
  - —air inlets and outlets
- Multiple benchmarks
  - —primarily Top500 (HPL)

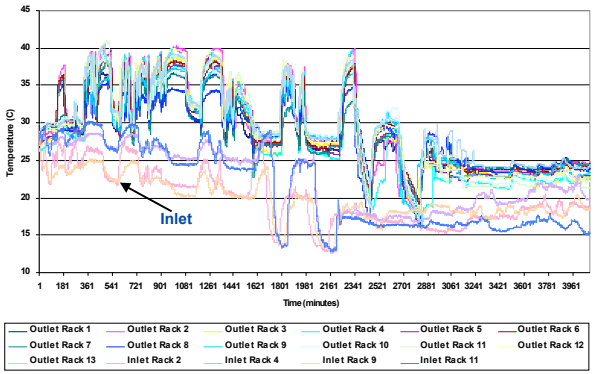
Mote Sensor Locations



Source: Shobana Ravi

LACSI

Large Cluster: Top 500 Benchmarking

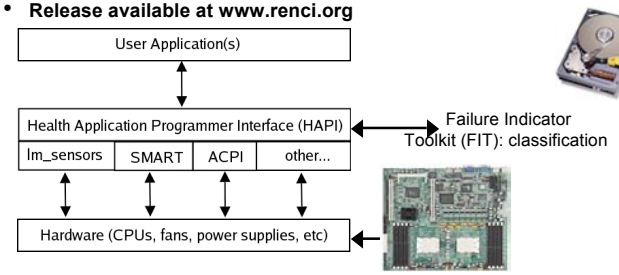


Source: Shobana Ravi

LACSI

UNC HAPI Implementation

- Health Application Programming Interface (HAPI)
  - —standard interface for health monitoring (by analogy with PAPI)
  - —ACPI (Advanced Configuration and Power Management)
  - —SMART (Self Monitoring, Analysis and Reporting Technology)
- Release available at www.renci.org

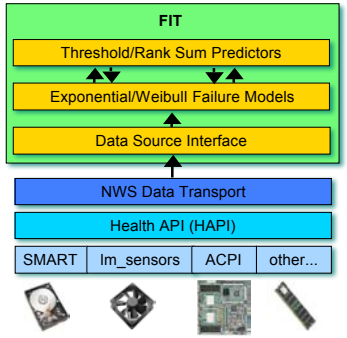


Source: Mark Reed/Kevin Gamiel

LACSI

Failure Indicator Toolkit (FIT)

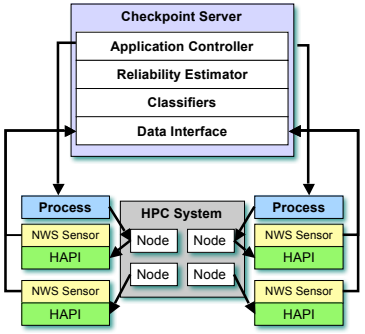
- **Concept**
  - —measure failure indicators
    - – disks, networks, …
    - – memory, motherboards
  - —predict likely failures
  - —adapt based on MTBF
    - – checkpoint frequency
    - – batch scheduling, …
- **Approach**
  - —standard data interfaces
  - —statistical classifiers
    - – failure prediction
  - —application controller
    - – adaptation



Source: Cory Quammen

LACSI

FIT Adaptive Checkpointing



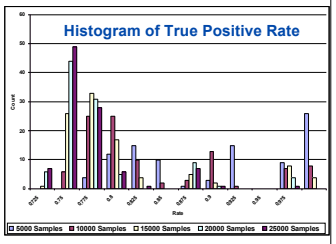
- **Checkpointing frequency**
  - — application driven
    - – susceptibility to faults
  - — reliability driven
    - – application needs
    - – system capabilities
- ***Adaptive checkpointing***
  - — *FIT MTBF estimate*
  - — *application controller*
- ***Experiments beginning …***

Source: Cory Quammen

LACSI

Failure Assessment Experiments

- **Disk data (from Murray et al)**
  - —177 good disks (tested at manufacturer)
  - —191 failed disks (customer returns)
  - —64 attributes (55 usable)
  - —observations every two hours
    - – up to 300 observations/disk
- **Assessment approach**
  - —randomly sample the population
    - – all observations from good disks
  - —determine min/max of attributes, e.g.,
    - – read head flying height (min)
    - – write errors (max)
  - —test each good and bad disk
    - – violation of threshold definitions
- **Preliminary results**
  - —71% accurate prediction
    - – with no false positives

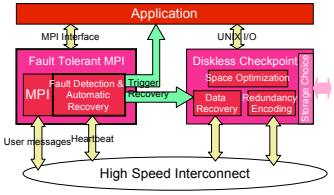


Source: Cory Quammen

LACSI

Large Scale Adaptation Examples

- ***Batch queue selection***
  - —*application fault sensitivity*
  - —*predicted partition reliability*
  - —*power/temperature constraints*
- **Checkpoint frequency**
  - —application fault sensitivity
  - —predicted partition reliability
- **Redundancy application**
  - —spare nodes for reliable execution
- **Power aware code optimization**
  - —tuning for power/performance/reliability
- **OS suicide hotline**
  - —adaptive personality management



LACSI

Job Scheduling Policies and Power

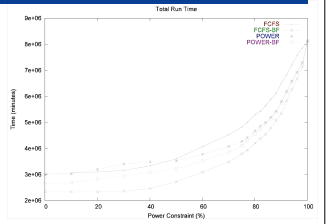
- Today, batch scheduling is largely power oblivious
  - —utilization and delay metrics dominate
  - —predominantly First Come First Serve (FCFS)
    - backfilling to improve utilization
- Power and temperature implications
  - —temperature transients lag job completion
    - cooling costs
  - —power budgets are increasingly important
    - fluctuating demands on power infrastructure
- Goals
  - —bound total power consumption
  - —minimize utilization and delay impact

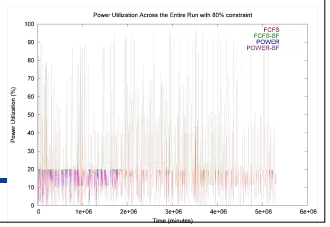
Source: Shobana Ravi

LACSI

Very Preliminary Evaluation

- LANL CM-5 workload
  - —122,055 jobs on 1024 nodes
  - —24 month period
- POWER
  - —scheduled ranked on power
- POWER-BF
  - —scheduled ranked on power
  - —backfilling ranked on power
- FCFS
  - —scheduled ranked on submit time
- FCFS-BF
  - —scheduled ranked on submit time
  - —backfilling ranked on submit time





Source: Shobana Ravi

LACSI Impacts

- Market forces and laboratory needs
  - —multicore chips and massive parallelism
    - capability and capacity systems
  - —power budgets ($) and thermal stress
    - economics and reliability
- Tools and systems haven't kept pace
  - —scale, complexity, reliability and adaptation
- *Making large systems more usable (our focus)*
  - —*scale, measurement and reliability*
  - —*power management and cooling*
  - —*prediction and adaptation*
- Federal policy initiatives
  - —June 2005 PITAC computational science report (chair)
    - "Computational Science: Ensuring America's Competitiveness"
  - —Computing Research Association (CRA) (chair, board of directors)
    - Innovate America partnership

LACSI