Correlated Resource Models of Internet End Hosts

Eric Heien

Derrick Kondo

David Anderson

Outline

- Problem and Objective
- Resource Overview
- Resource Model
  - Cores
  - Memory
  - Computation Speed
  - Disk Space
- Model Validation
- Comparison to other models

Problem

- Need resource models of Internet hosts
  - Scheduling algorithms
  - Application/system design
- Surprisingly few models for Internet hosts
  - Some work on Grid or cluster resources [1]
  - Numerous papers on Internet network architecture [2][3]
  - Some benchmark programs available, but data is not public, support only Windows, oriented towards game performance [4][5]

[1] Kee, et. al. "Realistic Modeling and Synthesis of Resources for Computational Grids", Supercomputing 2004
[2] Floyd, Koller "Internet Research Needs Better Models", 2003
[3] Faloutsos^3, "On power-law relationships of the Internet topology", 1999
[4] "PassMark," http://www.passmark.com/
[5] "LMBench - Tools For Performance Analysis," http://www.bitmover.com/lmbench

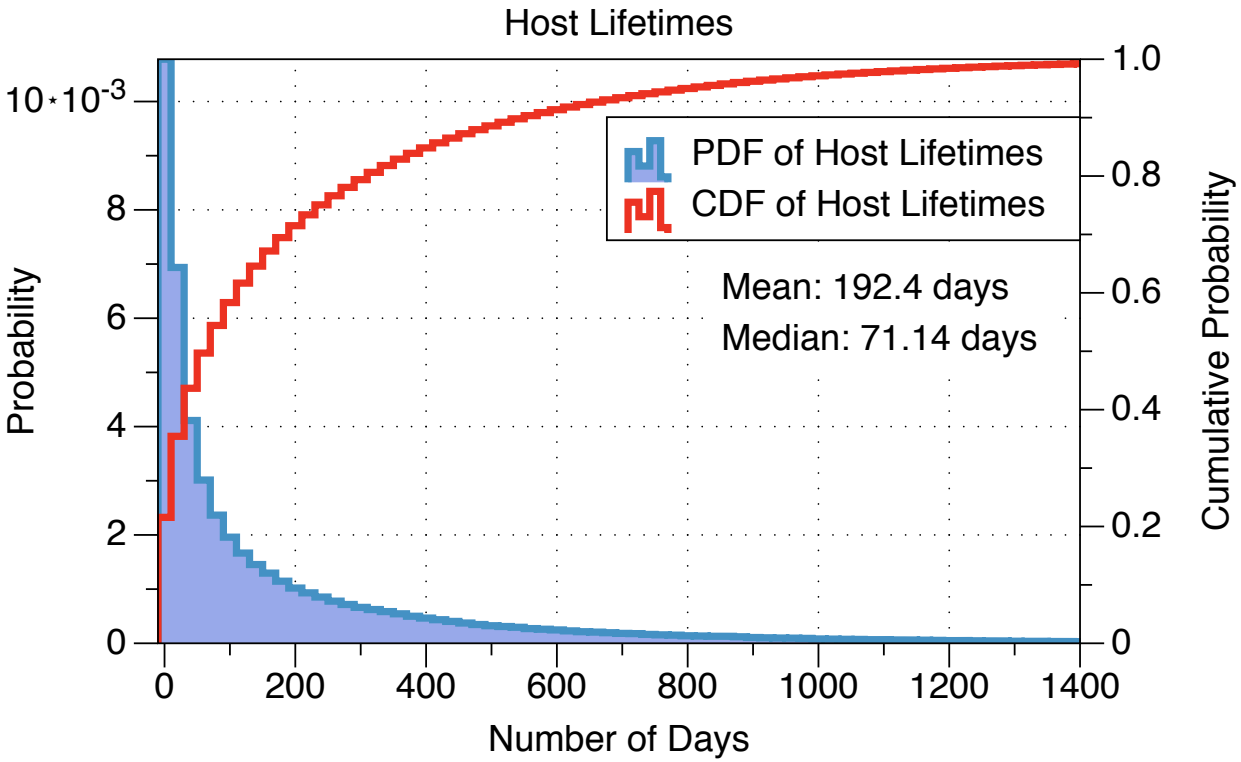
Objective

- Create resource model
  - Model # processors, processor speed, memory size, disk space
  - How do these resources change over time?
  - How are resources correlated?
- Validate resource model
  - How well does model match real data?
- Compare our model to other models
  - Is our model more accurate than others?

Methodology

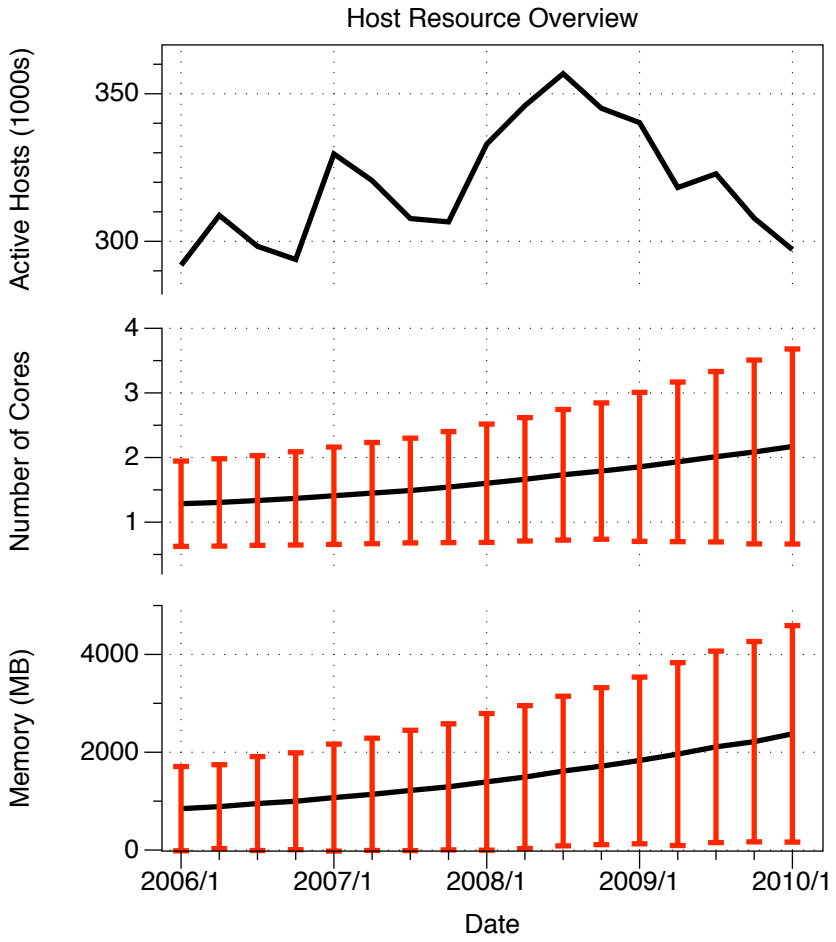
- Use BOINC (Berkeley Open Infrastructure for Network Computing) to measure hosts in SETI@home project
- Recorded resources of 2.7 million hosts from Jan 2006-Sep 2010
- Used Whetstone, Dhrystone to measure processor speed
- Remove questionable hosts from data set (e.g. > 128 processors, > 100GB memory, etc)

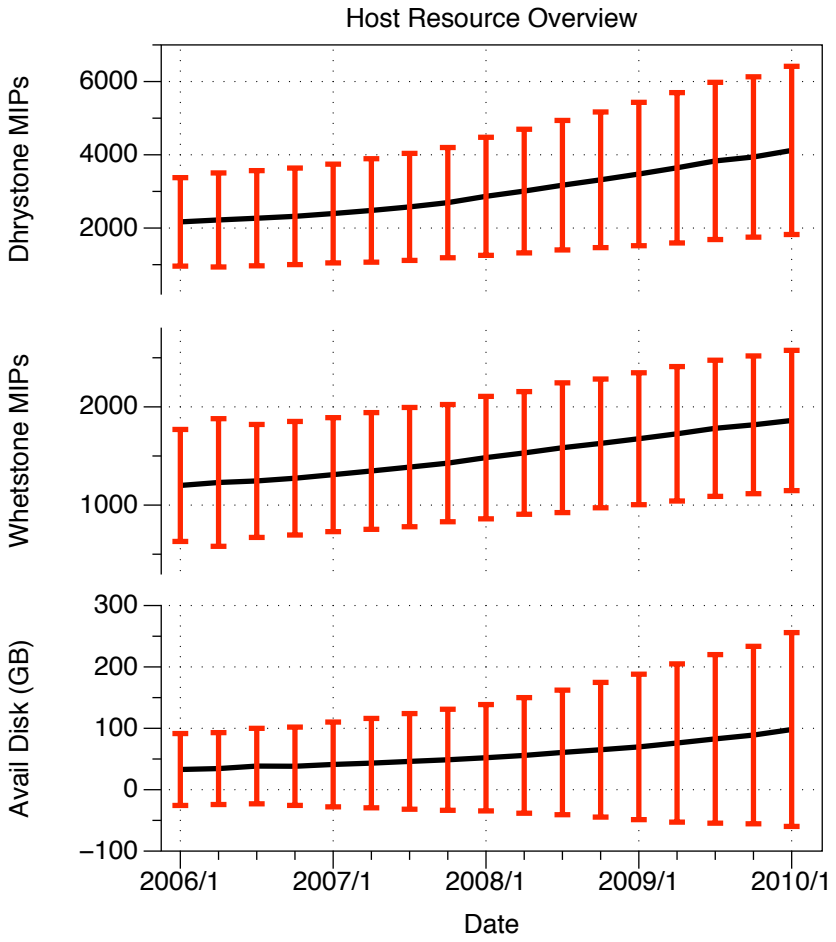
Resource Overview



- Well fit by a Weibull distribution (k = 0.58, λ = 135)
- Indicates decreasing dropout rate

Resource Overview





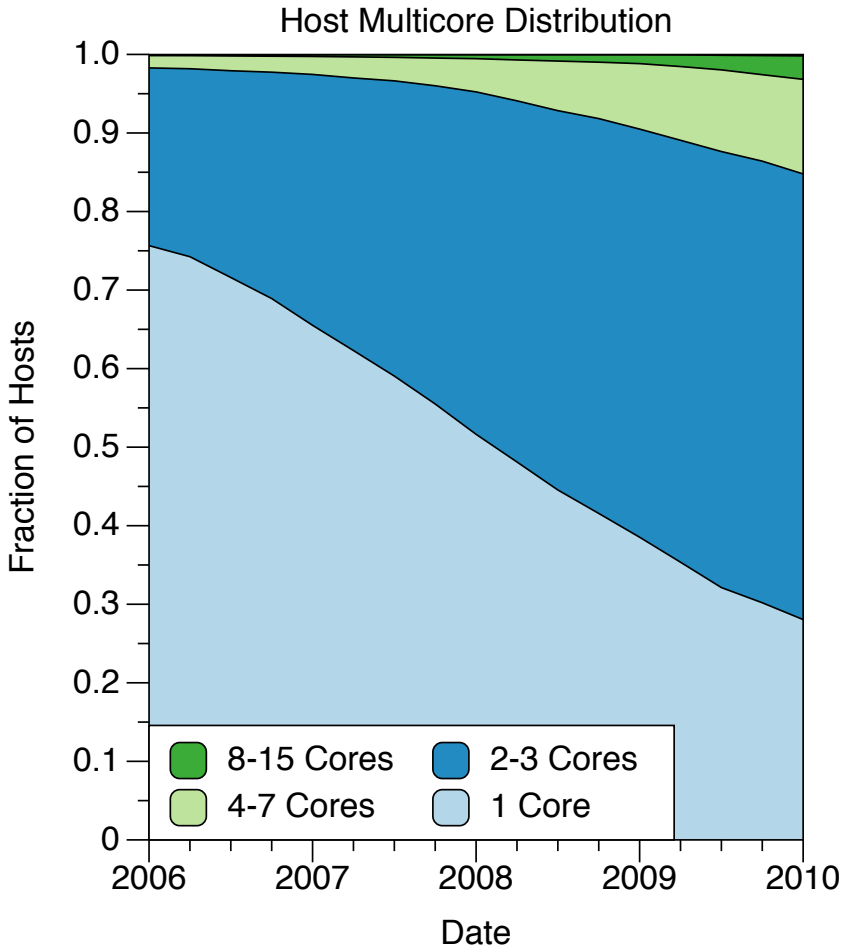
Resource Correlations

- Are resources correlated?
  - More processors indicates more memory? (yes)
  - More disk space indicates faster processor? (no)

	Procs	Memory	P-C-Mem	Whet	Dhry	Disk
Processors	1.00	0.61	-0.01	0.16	0.13	0.09
Memory	-	1.00	0.63	0.23	0.27	0.11
Per-Core-Mem	-	-	1.00	0.25	0.31	0.07
Whetstone	-	-	-	1.00	0.64	-0.02
Dhrystone	-	-	-	-	1.00	0.00
Disk Space	-	-	-	-	-	1.00

Resource Model – Cores

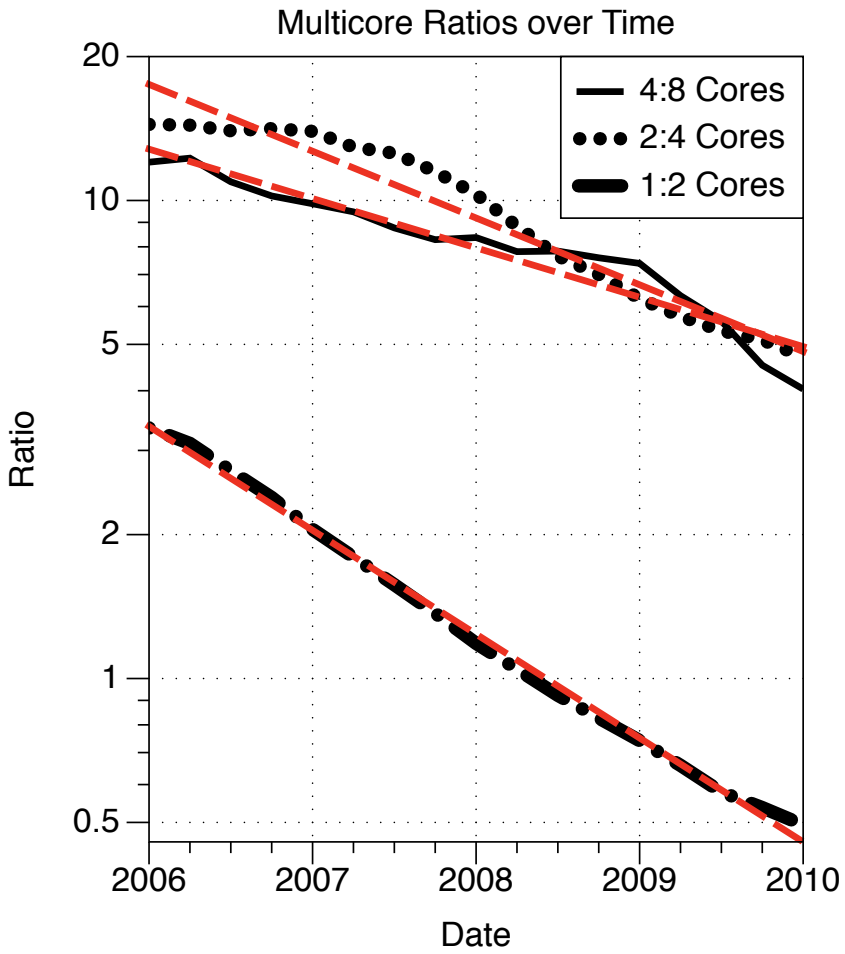
- Since active hosts fluctuate, model as fraction of total hosts
- Goes from mean of 1.28 to 2.17 cores (70% incr.)
- Poor fit for discretized truncated log-normal, log-gamma distributions
- Model # of cores as relative fractions



Resource Model – Cores

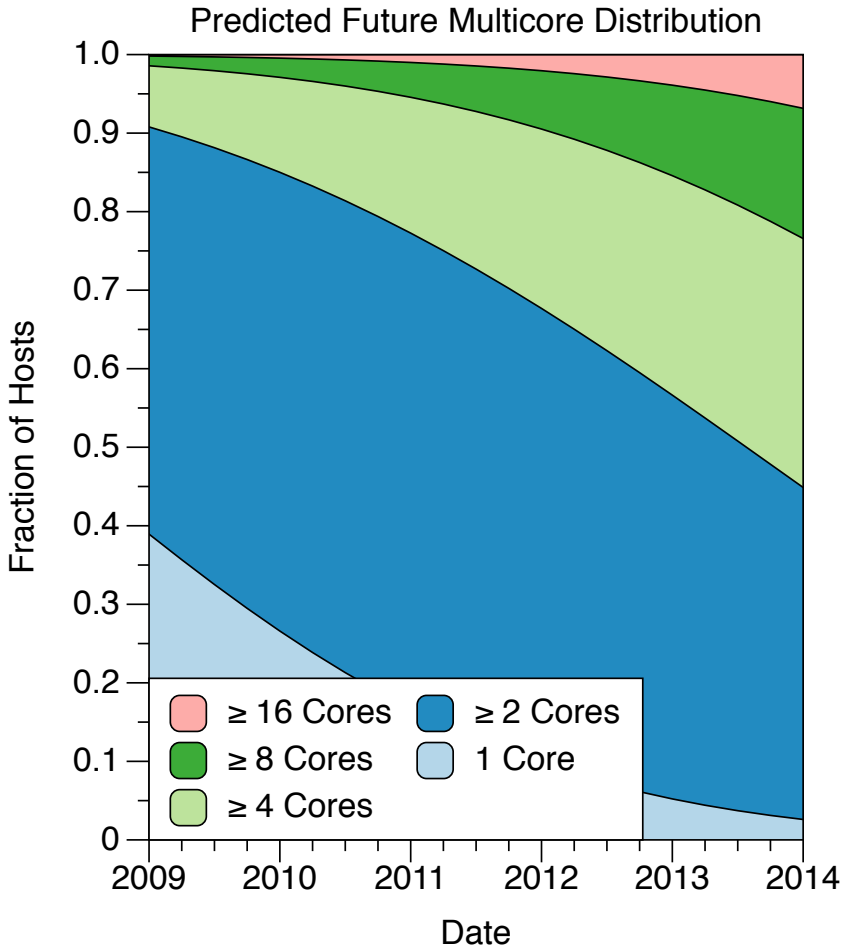
- Plot multicore ratios
  - (# hosts with N cores)/ (# hosts with 2N cores)
- Will be high initially
  - Most hosts have fewer cores
- Over time will decrease
  - Hosts with more cores become common
  - Older hosts drop out
- Well fit (r > 0.95) by an exponential curve

  Ratio = a*exp(b*(Date-2006))



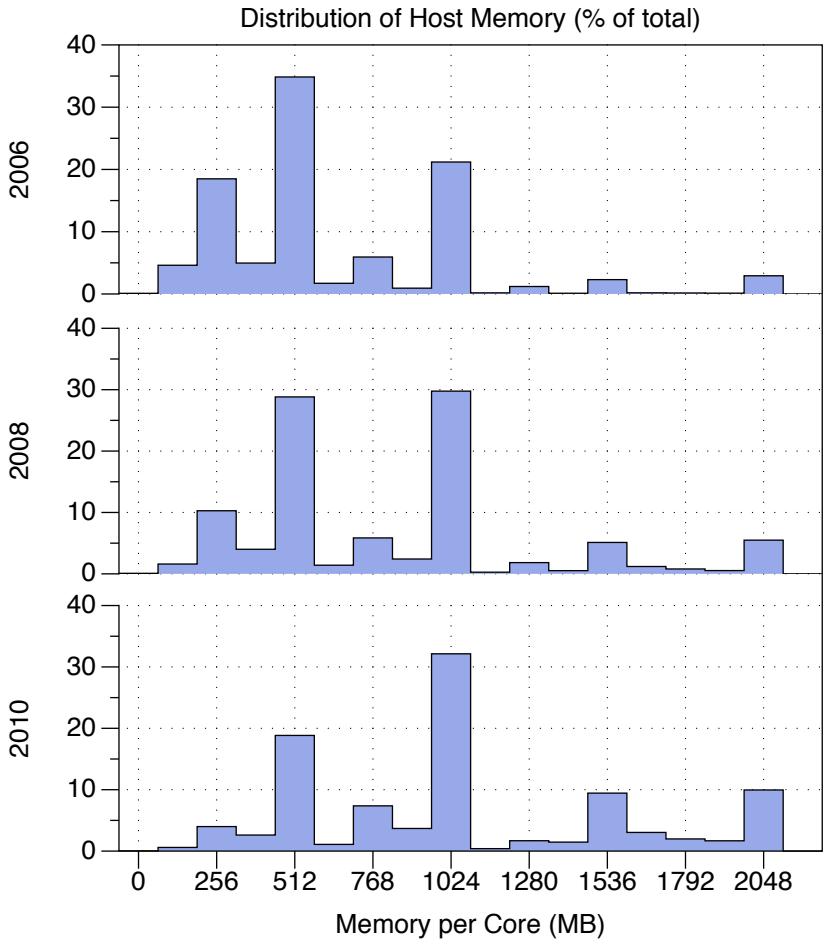
Resource Model – Cores

- Allows for prediction of future host composition
- By 2014 model predicts:
  - Mean of 4.6 cores per host, median of 4 cores
  - Single core hosts about 2.6% of total
  - Hosts with 2 cores most common (42.2%)
- Significantly different than simple extrapolation of mean value (3.7 cores)



Resource Model – Memory

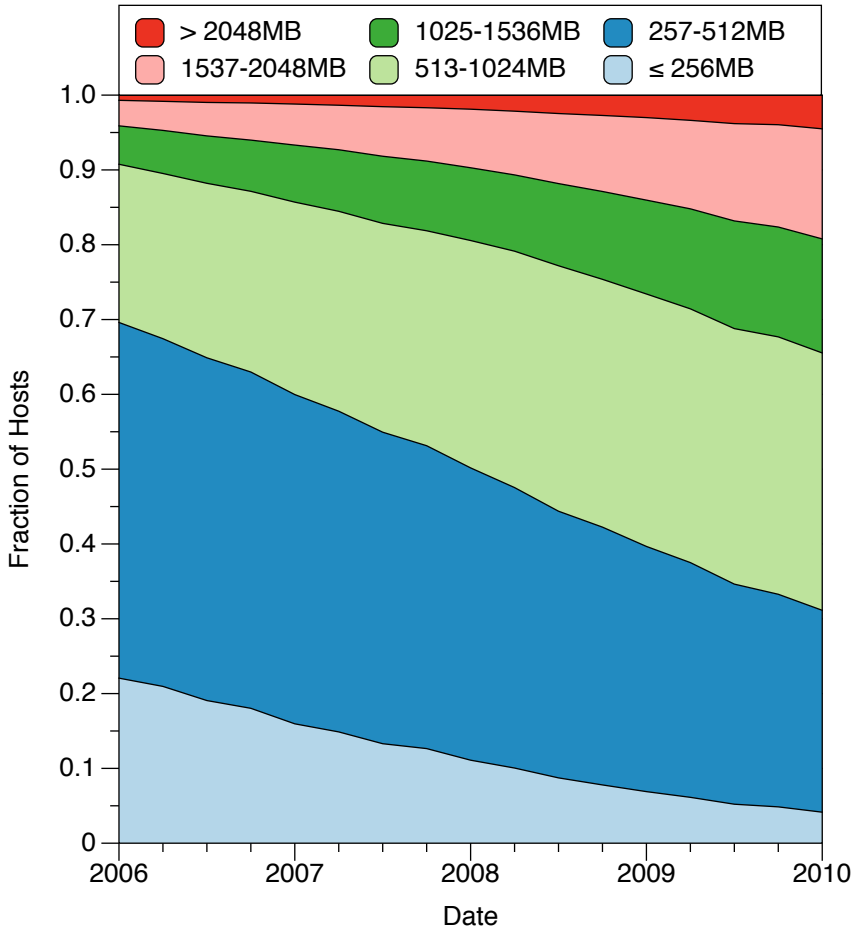
- From 2006 to 2010, mean host memory rises from 846MB to 2376MB (181% increase)
- Total memory is correlated with cores (p=0.6)
  - Per-core-memory is not correlated with cores
  - We model per-core-memory as an independent variable
- Highly irregular distribution
  - Poor fit to log-normal or log-gamma



Resource Model – Memory

- Similar to CPUs, plot ratios of memory
- Again, well fit (r > 0.97) by an exponential curve

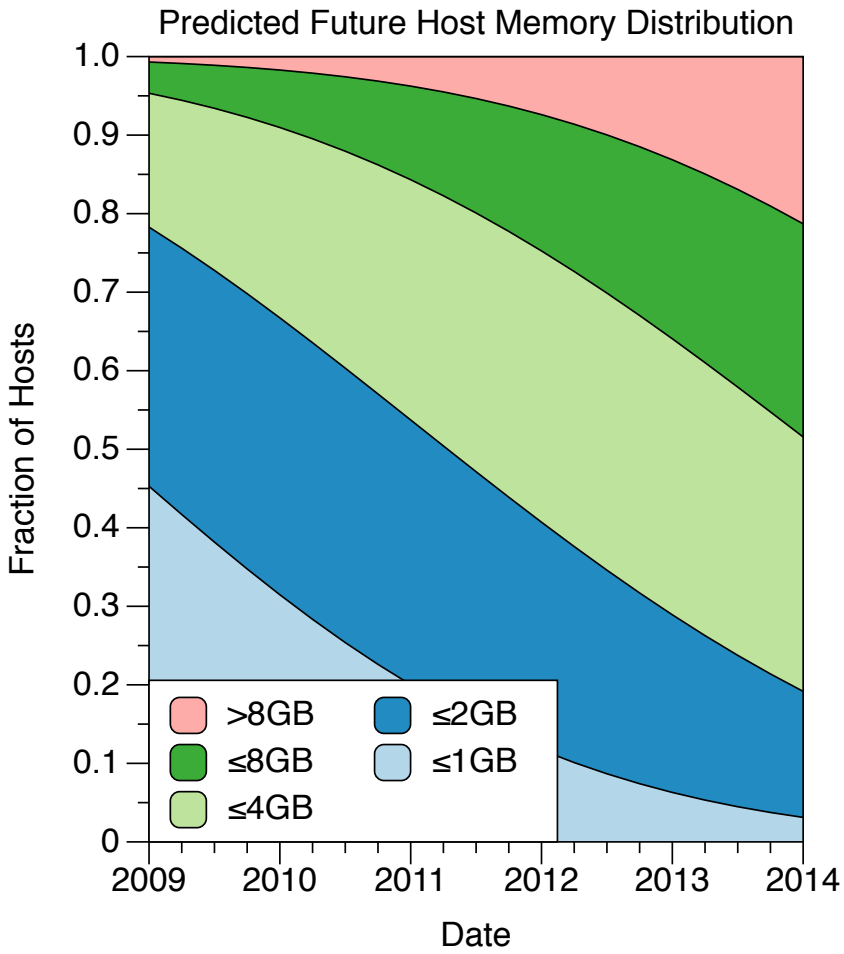
  Ratio = a*exp(b*(Date-2006))
- Can increase model accuracy by adding more ratios (2GB:3GB, etc)
  - Increases model complexity
  - Current ratios represent 80% of data set



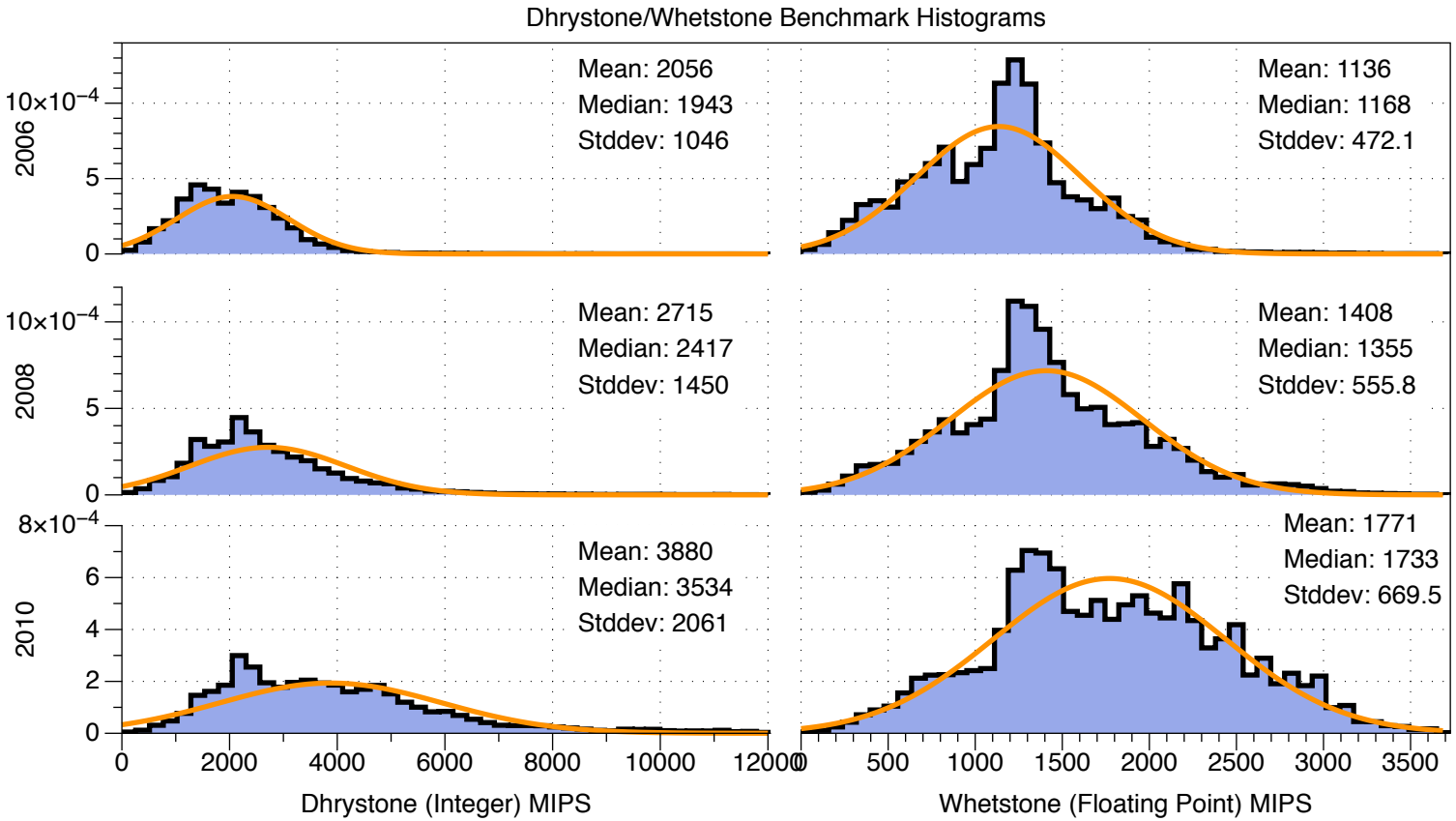
Per-Core-Memory Fractions

Resource Model – Memory

- Prediction for future host memory
- Combine core+per-core-memory models and extrapolate
- By 2014:
  - Mean of 6.8 GB per host, median of 4 GB
  - Few hosts (3%) with 1GB or less of RAM
  - Large fraction (21.3%) of hosts with more than 8 GB RAM



Resource Model – Speed



Somewhat irregular, but fit reasonably well by normal distribution

Resource Model – Speed

- Extrapolate values
  - Sample normal distribution with those characteristics
- Problem: speeds are correlated with each other and per-core-memory
- Solution: generate correlated distributions
  - Make correlation matrix (R)
  - Cholesky decomposition (U)
  - Multiply three normal distribution samples by U to get correlated values (V_C)

Correlation Matrix

$$R = \begin{bmatrix} 1 & 0.250 & 0.306 \\ 0.250 & 1 & 0.639 \\ 0.306 & 0.639 & 1 \end{bmatrix}$$

Cholesky Decomposition

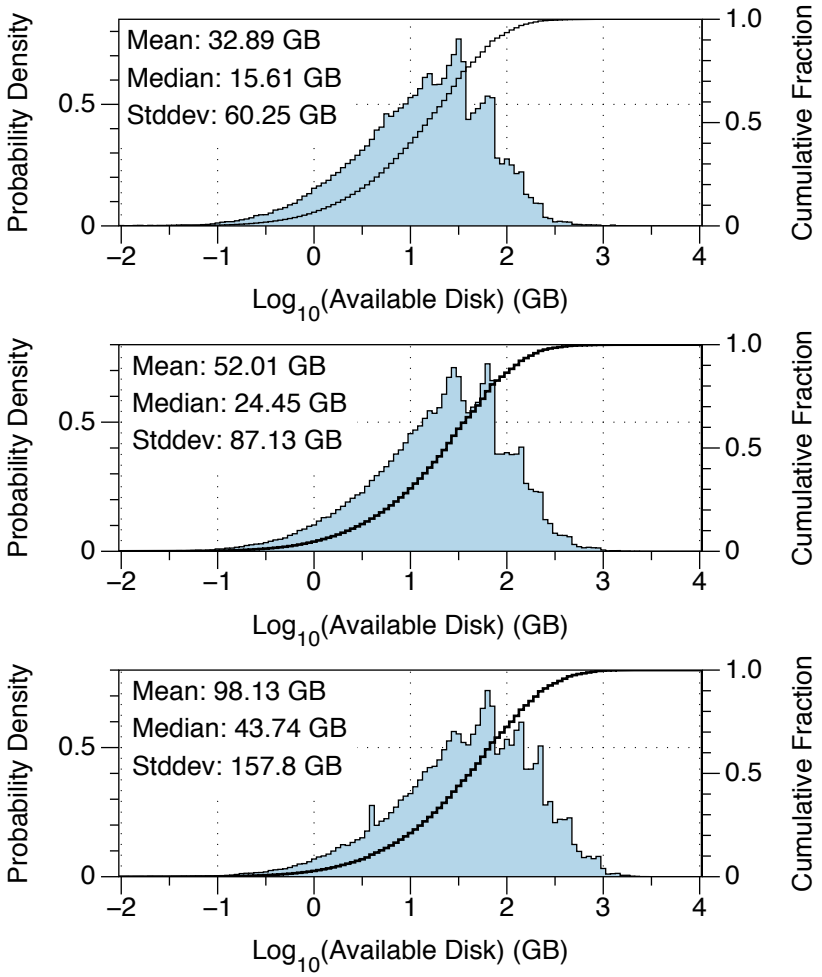
$$U = \begin{bmatrix} 1 & 0 & 0 \\ 0.250 & 0.968 & 0 \\ 0.306 & 0.581 & 0.754 \end{bmatrix}$$

$$V = \begin{bmatrix} Norm(0,1) \\ Norm(0,1) \\ Norm(0,1) \end{bmatrix}$$

$$V_C = UV$$

Resource Model – Disk Space

- Disk space is uncorrelated with other resources
  - Can generate independently
- Best fit to a log normal distribution
  - Log gamma also reasonable
- Extrapolate parameters and generate by sampling distribution



Resource Model – Summary

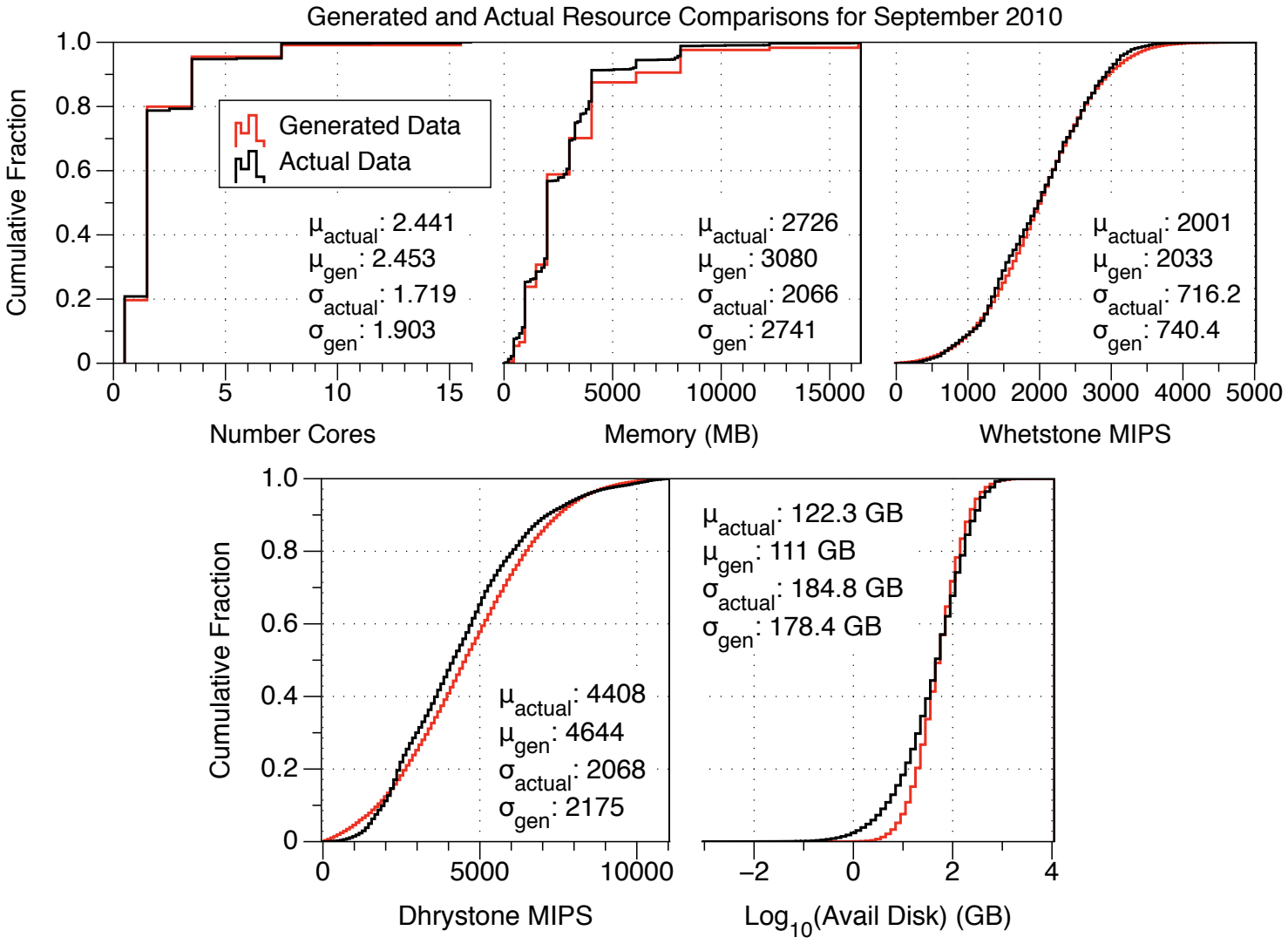
Resource	Value	Method	a	b
Cores	1:2 Core	Relative Ratio	3.4	-0.50
	2:4 Core	Relative Ratio	17.5	-0.32
	4:8 Core	Relative Ratio	12.8	-0.24
Per-Core-Mem	256MB:512MB	Relative Ratio	0.6	-0.25
	512MB:768MB	Relative Ratio	4.9	-0.13
	768MB:1GB	Relative Ratio	0.4	-0.17
	1GB:1.5GB	Relative Ratio	4.0	-0.14
	1.5GB:2GB	Relative Ratio	1.5	-0.09
	2GB:4GB	Relative Ratio	5.0	-0.10
Dhrystone	Mean (MIPS)	Correlated Normal Distribution	2064	0.17
	Variance	Correlated Normal Distribution	1.38e6	0.33
Whetstone	Mean (MIPS)	Correlated Normal Distribution	1179	0.12
	Variance	Correlated Normal Distribution	3.24e5	0.11
Disk Space	Mean (GB)	Lognormal Dist.	31.59	0.27
	Variance	Lognormal Dist.	2890	0.52

Model Validation

Actual Data	Procs	Memory	P-P-Mem	Whet	Dhry	Disk
Processors	1.00	0.61	-0.01	0.16	0.13	0.09
Memory	-	1.00	0.63	0.23	0.27	0.11
Per-Proc-Mem	-	-	1.00	0.25	0.31	0.07
Whetstone	-	-	-	1.00	0.64	-0.02
Dhrystone	-	-	-	-	1.00	0.00
Disk Space	-	-	-	-	-	1.00

Model	Procs	Memory	P-P-Mem	Whet	Dhry	Disk
Processors	1.00	0.73	0.01	0.00	0.01	0.00
Memory	-	1.00	0.54	0.16	0.14	0.00
Per-Proc-Mem	-	-	1.00	0.31	0.25	0.00
Whetstone	-	-	-	1.00	0.51	0.00
Dhrystone	-	-	-	-	1.00	0.00
Disk Space	-	-	-	-	-	1.00

Model Validation



Model Comparison

- Run a simulation using real host data and generated host data
- Data contains memory size, disk space, etc
  - How to include these in simulation?
  - Use utility function based on classic Cobb-Douglas
- Utility (Y) of application A on host H with C cores, M memory, I Dhrystone performance, F Whetstone performance and D disk space is:

$$Y_A(H) = C_H^{\alpha} M_H^{\beta} I_H^{\gamma} F_H^{\delta} D_H^{\epsilon}$$

- α, β, γ, δ, ε indicate affinity of application for resource

Model Comparison

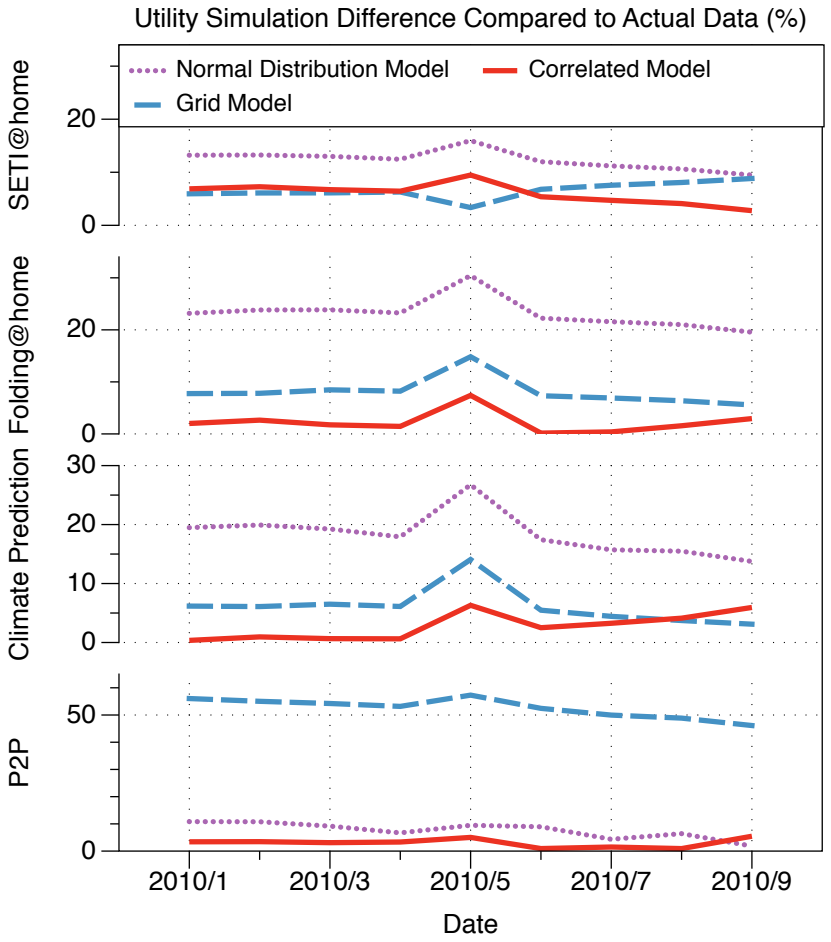
Application	Cores (α)	Memory (β)	Dhrystone (γ)	Whetstone (δ)	Disk (ε)
SETI@home	0.05	0.1	0.2	0.4	0.05
Folding@home	0.4	0.05	0.2	0.3	0.05
Climate Prediction	0.2	0.2	0.1	0.35	0.15
P2P	0.05	0.1	0.1	0.05	0.7

- Compared our model with two others:
  - "Normal distribution": Extrapolate means, variances and sample resources from uncorrelated normal distributions (lognormal for disk space)
  - "Grid Model": Kee et. al. [1] model of Grid resources updated with values from our data set
- Use greedy round robin algorithm to assign resources to application
  - Sum application utility and compare totals for different models

[1] Kee, et. al. "Realistic Modeling and Synthesis of Resources for Computational Grids", Supercomputing 2004

Model Comparison

- In general our model is closest to actual data
- Normal distribution model generally fares poorly
  - No resource correlation
- Grid model is slightly worse
  - Very low accuracy for disk space
  - Indicates difficulty of predicting resources based purely on hardware trends



Summary

- Created and validated model of Internet connected hosts
  - Captures cores, memory, speed and disk space
  - Captures correlations between these
  - Captures change over time
- Compared our model to other models
- Future work
  - Include GPU characteristics in the model
  - Investigate if resource correlations change over time
  - Investigate correlations between resources and location
  - Additional resource (cache, swap space, network, etc)