

The Heavy Tail: Challenges in Grid Marketplace Design



Kartik Hosanagar
Wharton School, UPenn
Ramayya Krishnan
Heinz School, CMU

Outline

- Examples of hosted and distributed computing services
- Grid Computing markets
- Heavy Tailed Demand in Grids
- Implications of Heavy tailed Demand
 - Estimating Resource Requirements
 - Pricing
 - Capacity Planning
- Conclusions

Key Trends in Distributed Computing

- Service Oriented Computing
 - Rapid low-cost composition of distributed resources
 - Service provider, service registry and service consumer
- Software as a service
 - Example: Salesforce.com
 - Salesforce (service provider) hosted on Sun's data centers (resource provider)
- Distributed Content Delivery
 - Example: Akamai

Salesforce.com

- On demand CRM
 - Appexchange API platform with Eclipse support
 - Permits integration of salesforce with other hosted services like Google Maps
 - Potential to create Mashups
- Software runs on multiple data centers
- Pricing
 - Pay per seat
 - \$65/user/month or \$699/5 users/year
 - Subscription model with no metering
 - Some contracts have SLA's related to uptime

Monitoring Service Levels

Salesforce.com: System Status - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Google

Back

Search Favorites

Links Customize Links Free Hotmail Windows Windows Marketplace Windows Media Cannot find server

Address http://trust.salesforce.com/ Go

Sunday May 14, 2006 | 7:55 am PDT

Service System	AP	EMEA	NA1	SSL
Status	✓	✓	✓	✓

All Systems Operational
No issues reported.

Service Performance History ↑

Date	Number of Transactions	Avg. Speed* (milliseconds)	System Status			
			AP	EMEA	NA1	SSL
05/13/06	13,072,323	202	✓	✓	✓	✓
05/12/06	37,104,265	265	✓	✓i	✓i	✓i
05/11/06	41,965,614	260	✓	✓	✓	✓
05/10/06	42,144,604	258	✓	✓	✓	✓
05/09/06	43,300,628	256	✓	✓	✓	✓
05/08/06	43,407,689	255	✓	✓	✓	✓
05/07/06	15,100,735	202	✓	✓	✓	✓
05/06/06	13,205,336	201	✓	✓	✓	✓
05/05/06	37,272,028	253	✓	✓	✓	✓
05/04/06	40,454,496	277	✓	✓	✓	✓
05/03/06	43,529,198	250	✓	✓	✓	✓
05/02/06	43,863,643	251	✓	✓	✓	✓

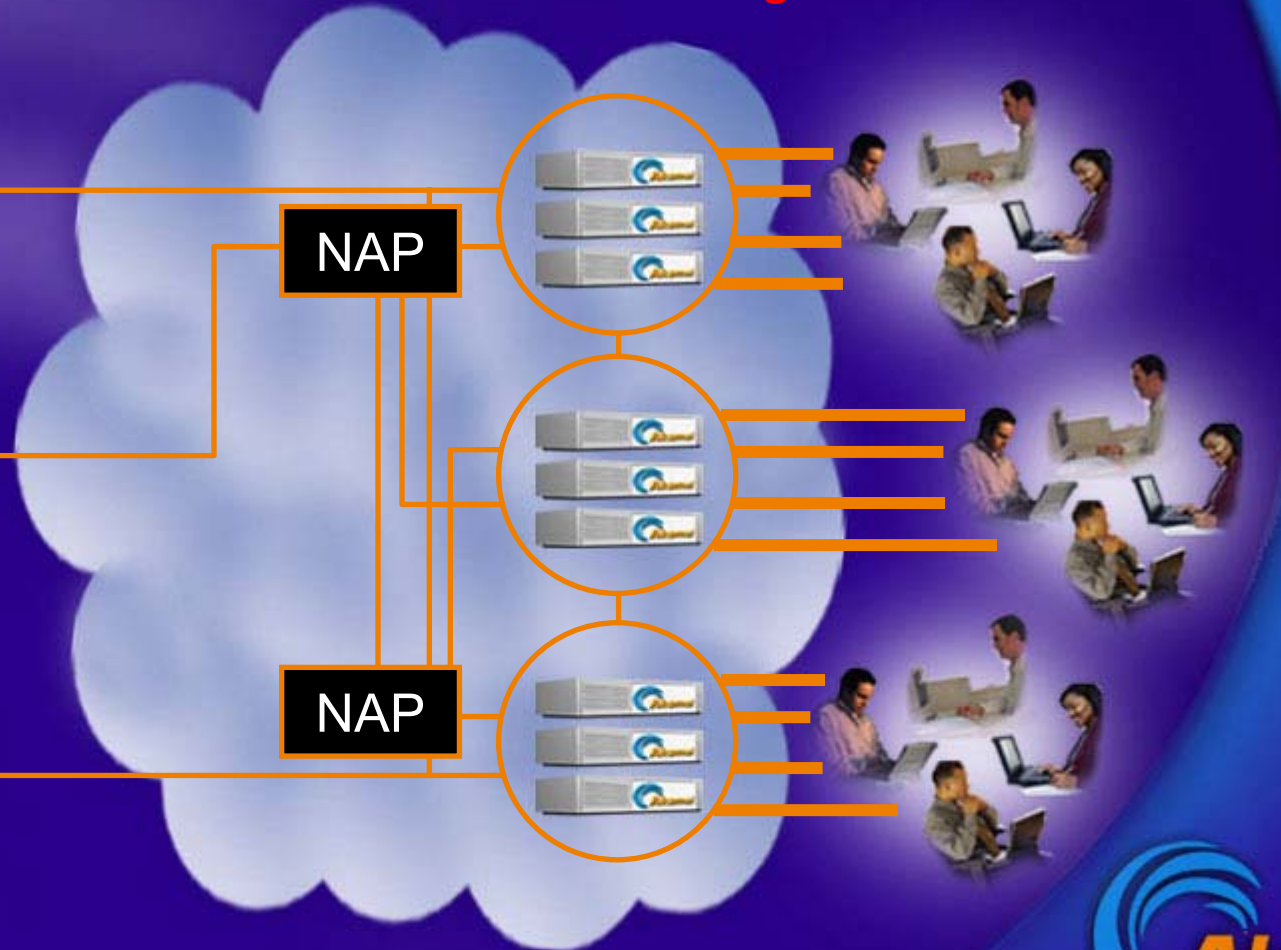
Done Internet

Akamai's Platform for Delivering Content and Applications

Content Providers

Akamai Servers at Network Edge

End Users



Akamai

- Content delivery as a service
 - Edge delivery
- Content resides on Akamai's distributed platform (servers cohosted with ISPs)
- Evolution
 - Free flow: Deliver bandwidth intensive objects such as images and video
 - Edgesuite: Entire site assembled and delivered from the edge
- Demand: Significant burstiness (heavy-tailed)
- Pricing
 - \$/mbps
 - 95th percentile of usage as measured by Akamai over five minute intervals for each CP Code ("95/5").
 - Any monthly usage in excess of two times the 95/5 for particular Property (CP Code) will be charged at the rate of \$[**] per GB delivered.
 - See <http://contracts.onecle.com/akamai/microsoft.svc.2004.09.01.shtml> for structure

Grid Computing

- Shared use of loosely coupled and distributed IT resources across geographies
 - Typically, the resource is CPU cycles
 - What about software licenses?
- Examples:
 - Berkeley's SETI project
 - Intel's NetBatch system
 - Singapore national grid
- "Grid" Marketplaces
 - Providers and buyers of computing cycles can come together in a market-based setting
 - Providers include IBM (\$0.47 per CPU hour), Sun (compute \$1/CPU hour or storage \$1/GB-mo)

Sun's Grid

<u>Elements</u>	<u>Sun's Grid</u>
Industry Standard Server	<u>V20Z Opteron (2.4 GHz),</u> <u>V210 SPARC</u>
RAM per CPU	4 Gig
Cache storage per CPU	20 Gig
Operating System	<u>Solaris 10</u>
Is OS open source?	<u>Yes</u>
Is OS Protected by ALL* corporate patents?	Yes
Minimum Commitment	4 hrs.
Price per hour	\$1 US

The Intel NetBatch System

- *"The idea behind NetBatch is very simple. An engineer simply submits a simulation job, whatever he wants to run, with NetBatch, hits enter, it then goes off to a broker that then assigns that task to one of the many resources that are part of the NetBatch queue anywhere across the Intel computing network. And as you see in this picture, it could be Israel, could be Chandler, Santa Clara, Folsom, any of those potential sites become participants or recipients of that particular job that's being run."*

Heavy Tail in Intel's NetBatch System

- **Systime: System (storage (I/O) time in seconds used by a Job**
- **Utime: User time used by a job in seconds**

	Mean	Std. Dev
Sys Time	76.62	818.51
User Time	3612.8	12909.22

Implications of the Heavy Tail

- Challenges in estimating resource requirements
 - For users
- Challenges in determining pricing schemes
 - For Suppliers
- Capacity Planning
 - For Suppliers

Estimating Resource Requirements

- Most scheduling or resource allocation techniques assume resource requirements can be estimated a priori
 - Scheduling matches jobs to resources based on requirements and availability
- The heavy tail can imply difficulties in predicting resource requirements
 - Users provide inaccurate estimates (Mu'alem and Feitelson 2001; Tsafrir et al 2005)
 - Not tied to incentives (backfilling, kill long jobs) but inherent difficulty

Suggested Directions

- Decision support tools for users can play an important role through data analysis and predictive modeling
 - Jobs in Net batch are part of Tasks
 - Analysis can be done at the task level and the user level

Stime,Utime,MaxRSS,ExitCode,Qslot,User,JobID,Task,StartTime,FinishTime,Workstation,osver,cpunum,cpuspeed,memory,swap,mode				
0.817,505.604,67384.0,0,1000,1001,74694598,1002,05/01/2006 10:18:57,05/01/2006 10:27:41,1003,1004,1,2200,2048,4096,-1				
0.745,432.301,67548.0,0,1000,1001,74694599,1002,05/01/2006 10:18:59,05/01/2006 10:26:24,1005,1004,2,3200,4096,8197,-1				

Mean and Std. Dev. Of Utime By Task

Task	StdDev	Mean	numentries
1002	477.48	599.63	648
1020	6070.12	8365.85	19
1026	5916.41	10188.76	53
1039	5119.45	6823.10	18
1062	7092.60	10020.79	57
1091	6619.27	12158.00	23
1182	5930.16	11584.79	25
1184	6139.44	7875.03	56
1366	26757.51	109872.02	92
1608	6271.09	12903.45	55
1652	570.05	161.34	77
1908	59.53	183.15	32
1982	6938.02	12018.19	56

Notice the considerable reduction in the variance

Mean and Std. Dev. Of Utime By User

User	StdDev	Mean	numentries
1001	205.33	398.57	86899
1009	18524.36	5945.64	101
1019	7055.05	10711.79	905
1025	13741.23	9734.05	846
1030	12343.45	4479.61	614
1045	37871.88	23011.18	2648
1104	20954.52	18323.89	172
1194	477.57	402.80	234
1277	3845.09	9251.88	225
1365	55299.55	76631.85	132
1406	15606.54	10919.60	81
1434	17430.42	23997.75	189
1651	13226.12	19633.32	2242

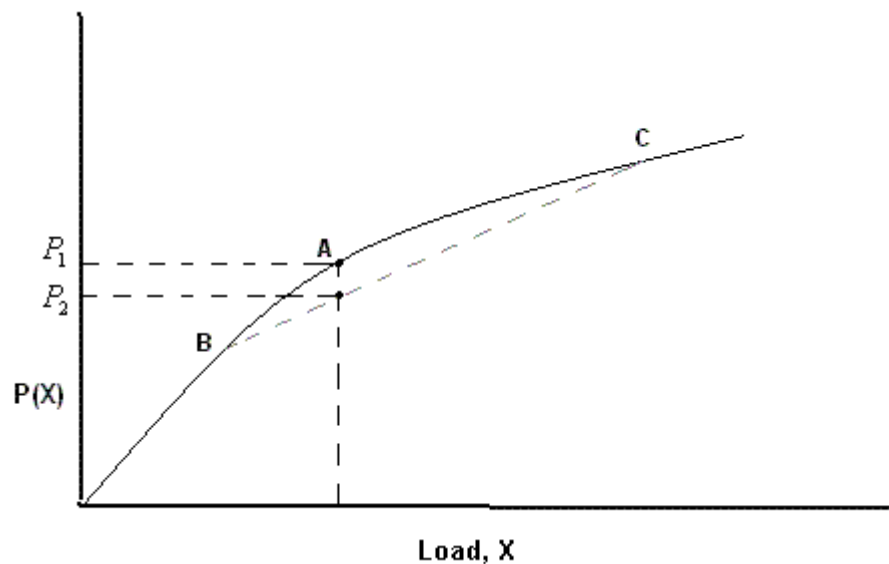
Notice the considerable reduction in the variance

Summary

- Predictive models to help users estimate resource requirements at the task and user level are likely feasible
- Use of these tools will have to be combined with mechanisms that are “strategy proof” that will induce truth telling behavior (see Parkes et al.)
 - How should these mechanisms accommodate uncertainty?

Pricing

- A job from the (heavy) tail imposes significant high cost on the system
 - Can a simple usage-based price (\$/CPU hr) account for this?
- Concave Pricing
 - Pros
 - Volume discounts
 - reflect economies of scale
 - Cons
 - Penalizes users with low variance ($P_1 > P_2$)



Pricing

□ Convex Price

- Penalizes users with high variance (with jobs in the tail)
- Quantity tax does not reflect economies of scale (volume discounts are desirable)

□ Suggested Directions

- Issue also faced by CDNs
- Percentile-based pricing can provide volume discounts and yet charge for high variance
- In contrast, Sun's price is \$1/CPU hr and IBM's price is \$0.47/CPU hr

Pricing Scheme (one-side and 'step' nonlinear pricing) cellular services

Plan #	Plan name	Fixed Fee	Free minutes	Overtime charge (<u>constant ratio</u>)
1	Pioneer	350	350	3/min
2	Option800	800	517	3/min
3	Option1100	1100	917	3/min
4	Option1500	1500	1217	3/min
5	Option2000	2000	2117	3/min

Comparisons with Cellular Usage and Pricing

Whole Sample					
	Mean	S.D.	Min	Max	N
Voice (Minute)	281.3	252.6	10.1	11845.3	59866
SMS (Message)	11.8	34.2	0	3106	
G1: Age <30					
	Mean	S.D.	Min	Max	N
Voice (Minute)	329.8	299.5	10.2	11845.3	22483
SMS (Message)	16.7	36.7	0	1558	
G2: 30<=Age <40					
	Mean	S.D.	Min	Max	N
Voice (Minute)	265.0965	223.455	10.08	6602.36	29520
SMS (Message)	9.232893	33.69502	0	3106	
G3: Age>=40					
	Mean	S.D.	Min	Max	N
Voice (Minute)	203.6786	168.1059	10.05	2916.9	7863
SMS (Message)	7.315656	26.84394	0	1079	

Average Usage Statistics 1 (across groups segmented by age)

Capacity Planning

- What are the resource elements that make up the grid?
 - CPU
 - RAM per CPU
 - Cache Storage per CPU
 - Software Licenses
 - Network capacity
 - ...
- What is the Service Provider's capacity planning problem?
 - How should resources be leased to meet stochastic demand?

Sun's Resource Grid

<u>Elements</u>	<u>Sun's Grid</u>
Industry Standard Server	<u>V20Z Opteron (2.4 GHz),</u> <u>V210 SPARC</u>
RAM per CPU	4 Gig
Cache storage per CPU	20 Gig
Operating System	<u>Solaris 10</u>
Is OS open source?	<u>Yes</u>
Is OS Protected by ALL* corporate patents?	Yes
Minimum Commitment	4 hrs.
Price per hour	\$1 US

Capacity Planning

- Set capacity to handle peak demand (excess capacity most of time) or something lower (low QoS at peak periods)?
- Pooling helps deal with variance
- Pricing scheme that accounts for variance (percentile based pricing)
 - Forces buyer to factor cost imposed (on others) by a burst before submitting jobs
 - Unlike CDN, buyer can control demand

Capacity Planning – Centralized vs. Decentralized

- In grids where resources are provisioned across different admin and geographic domains, locus of control is local
- Since perceived demand is local, decisions are locally optimal but don't take the global picture into account
- What is the "Price of anarchy" in this context?

Capacity Planning – a queuing framework

□ Cost of decentralization?

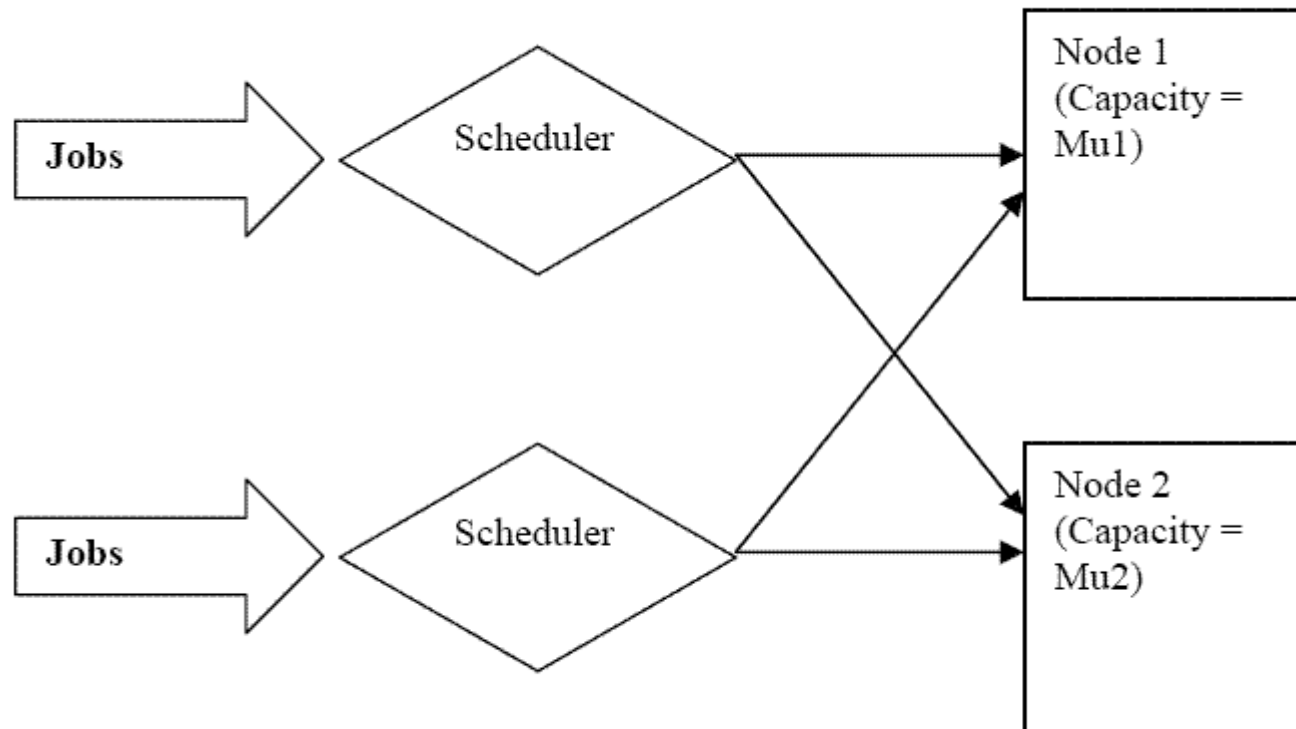


Figure 1. Queuing System for the Two Node Grid

Conclusions

- Predictive modeling taking variance in resource utilization
 - Has implications for customers and suppliers alike
- Resource allocation mechanism design with noisy estimates
- Capacity planning in the face of noisy demand estimates

Discussion/Conclusions

