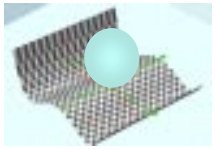
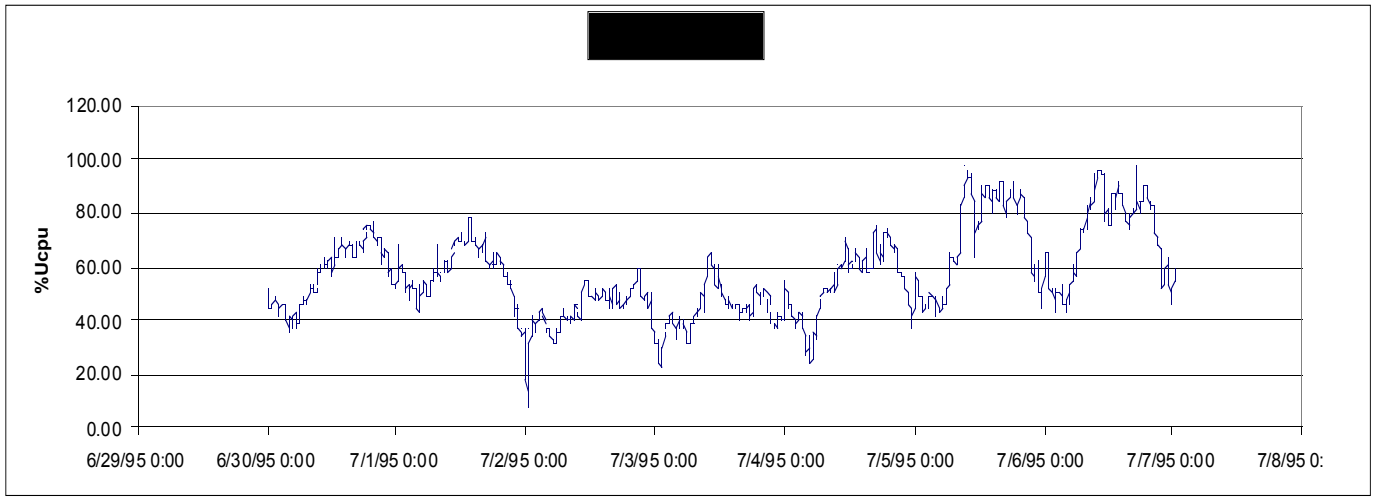
How to Assess Web Server Growth



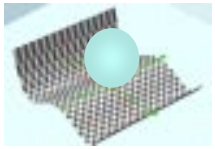
© 2000 Performance Dynamics Company™

Website Capacity Planning

- **Is capacity *planning* just a web oxymoron?**
- **Abundantly clear that capacity is needed**
- **But can the upgrade points be scheduled?**
- **We'll pull two rabbits out of the hat:**
  - **→ Short-term demand**
  - **→ Long-term growth**



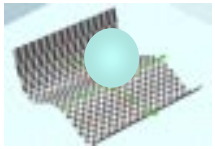
Database Server



© 2000 Performance Dynamics Company™

Statistical Models

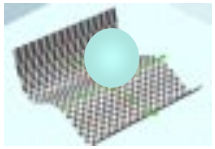
- **Useful for finding trends in raw performance data**
- **Trends can be used to predict growth**
- **Methods include:**
  - **→Regression**
    - **➢ time-independent data**
  - **→Time Series Analysis**
    - **➢ correlated in time**
  - **→ANOVA: ANalysis Of VAriance**
    - **➢ useful for quantifying the degree of interaction between otherwise random performance metrics**



© 2000 Performance Dynamics Company™

Why EXCEL?

- **Ubiquitous**
  - → Many desktops run MS Office
  - → Internet news groups for EXCEL/VBA **help**
- **Programmable**
  - → VBA, object-oriented, debugging, journaling, good prototyping tool
- **Data filtering**
  - → Need to internally select different data according to specified criteria
- **Statistical analysis functions (Regression, ANOVA, ARIMA)**
  - → Didn't know which ones would be needed to analyze this data
- **Charts**
  - → Plotting is integrated with data and has a VBA programmable interface
- **Printing**
  - → Single push-button
  - → Can optionally publish HTML pages
- **Limitations**
  - → I had not programmed in VBA before
  - → Dummy variables are a pain!
  - → Saving data and charts requires keeping separate workbooks

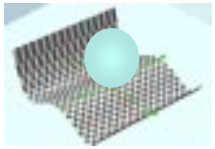


© 2000 Performance Dynamics Company™

Statistical Performance Toy

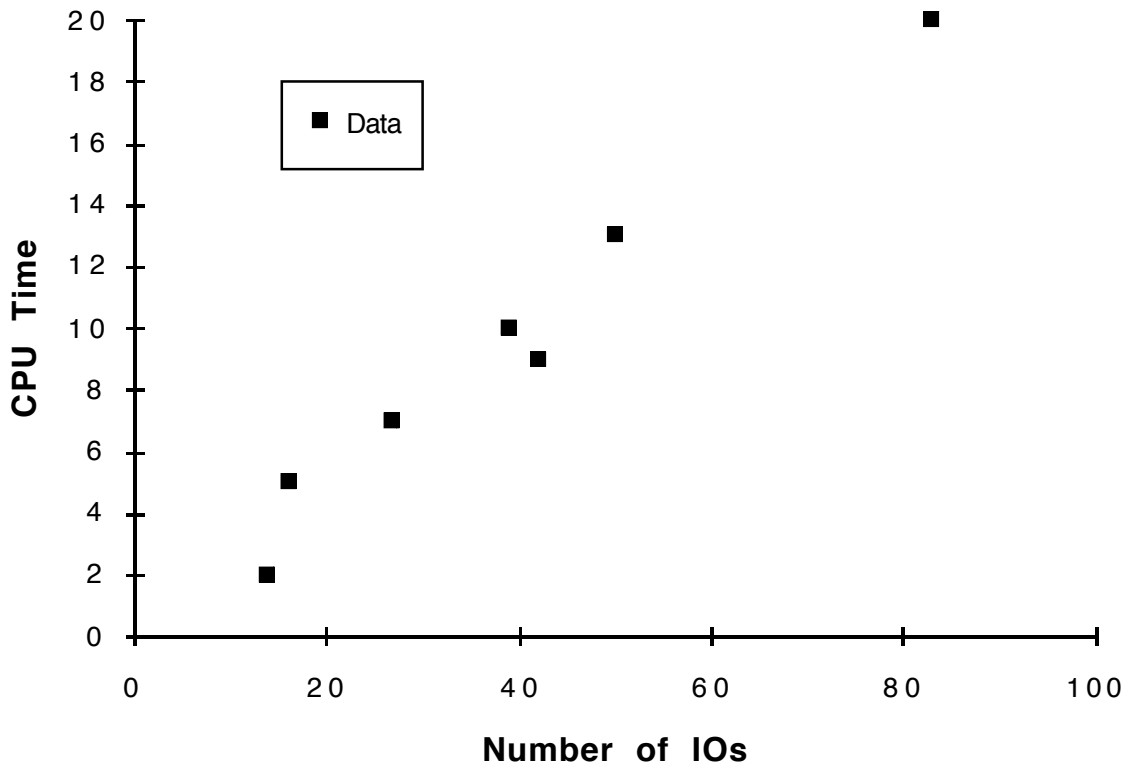
- **Collected performance data**
- **Total number of IOs issues**
- **Cumulative CPU time**

Total IOs	CPU Time
14	2
16	5
27	7
42	9
39	10
50	13
83	20

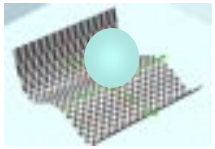


© 2000 Performance Dynamics Company™

Scatter Plot



Looks like a linear relationship, but is it?



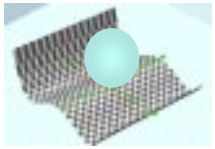
© 2000 Performance Dynamics Company™

EXCEL Regression Tool

Regression Statistics						
	Multiple R	0.98563213				
	R Square	0.9714707				
		0.96576484				
		1.08340978				
	Observations	7				
Analysis of Variance						
		df		*Mean Square*	*F*	*Significance F*
	Regression	1	199.845402	199.845402	170.258442	4.7161E-05
	Residual	5	5.86888379	1.17377676		
	Total	6	205.714286			
		Coefficients		*t Statistic*	*P-value*	*Lower 95%*	*Upper 95%*
	Intercept	-0.0082824	0.831105	-0.0099655	0.99237191	-2.1447023	2.12813755
	x1	0.24375637	0.01868107	13.0483118	1.2488E-05	0.19573524	0.2917775

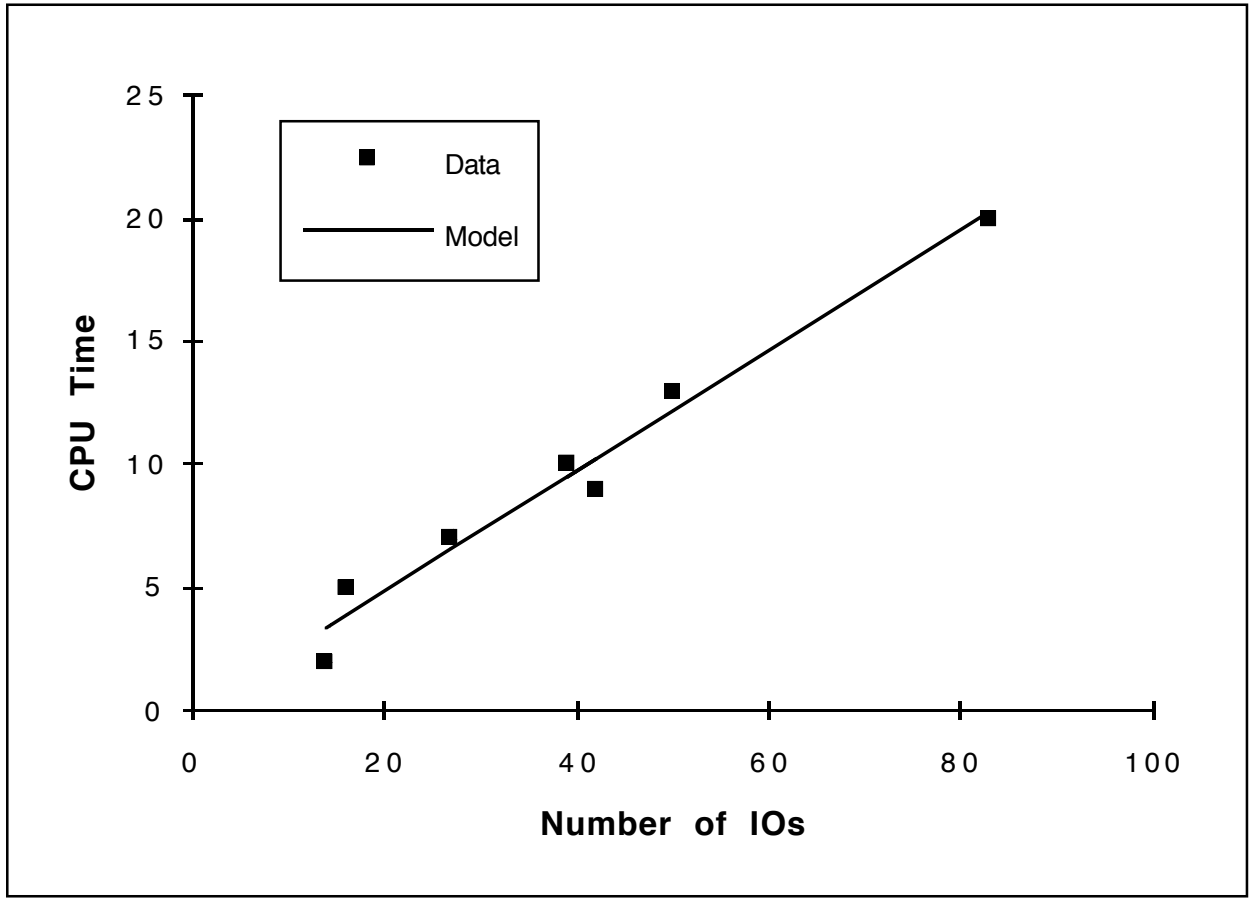
Linear model: y = mx + c

CPUtime = 0.24375637 (IOs) – 0.0082824

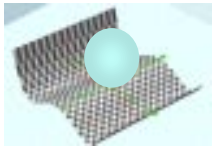


© 2000 Performance Dynamics Company™

Linear Regression Model

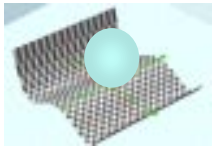


Use linear model to project IO and CPU capacity



© 2000 Performance Dynamics Company™

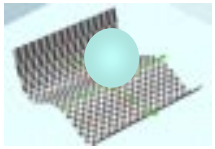
Multivariate Linear Regression



© 2000 Performance Dynamics Company™

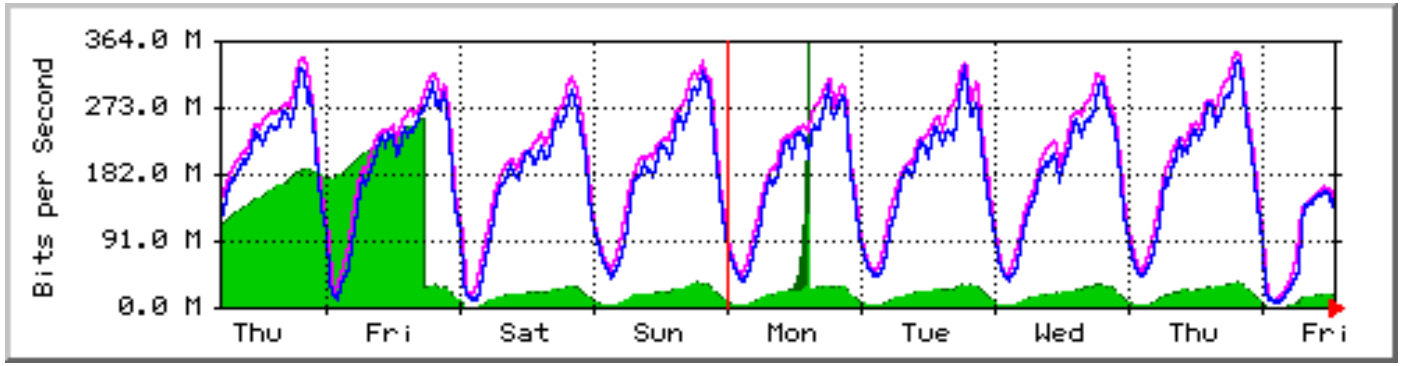
Web Server Weekly Data

- **Analyzed 10 weeks of data from 7/1/99 to 9/6/99**
- **Resorted to *SE Percolator* data (2 min samples)**
- **First 2 weeks (until 7/15) show Server > 160% busy**
  - **→At least 60% more capacity being demanded than was available**
  - **→Many upgrades (e.g., CPUs, ORACLE, Apps) from 7/15 to 7/31**
- **Only the last 5 weeks of data are stable enough to show consistent trending information**

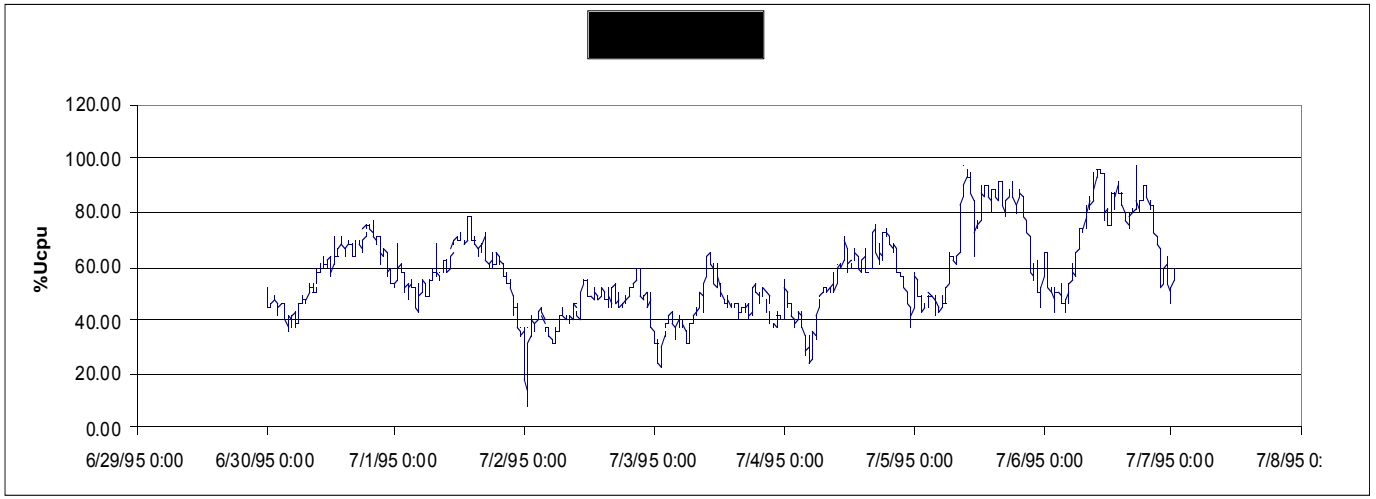


© 2000 Performance Dynamics Company™

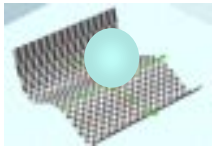
Daily Data



Network

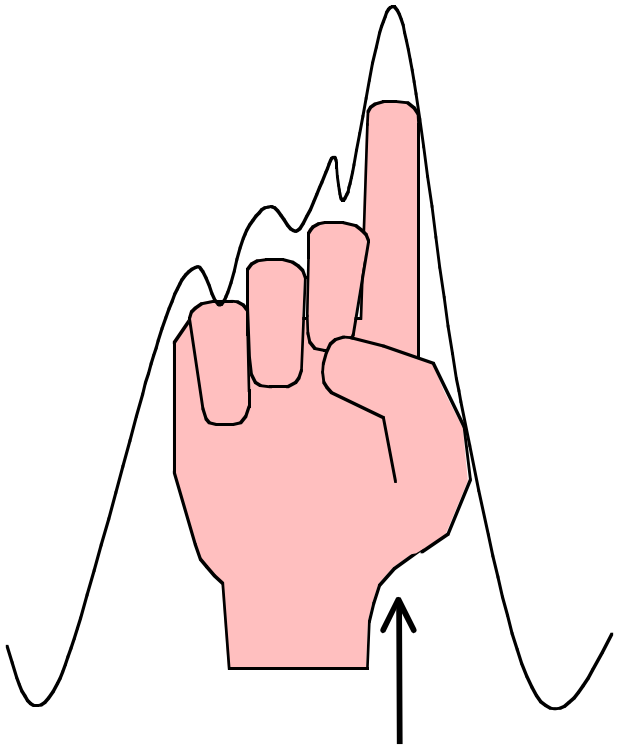


Database Server

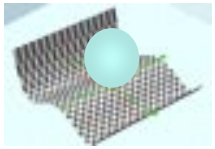


© 2000 Performance Dynamics Company™

Numero Uno Signature

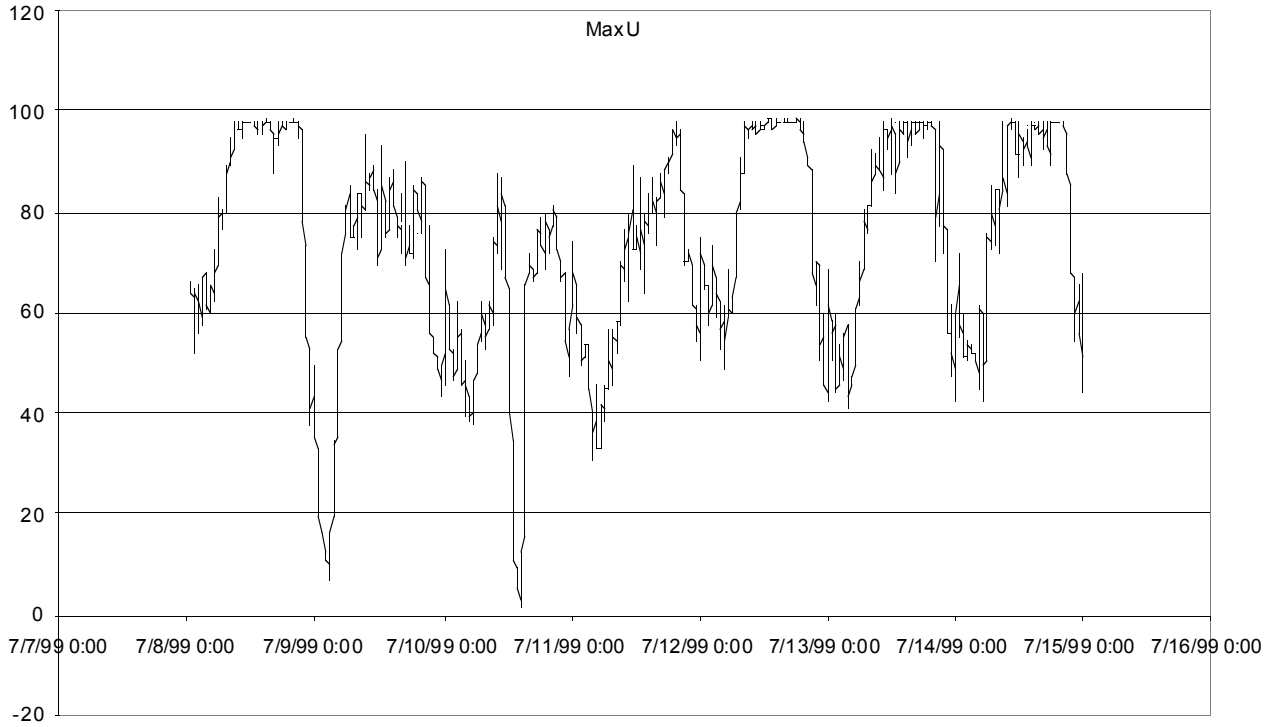


- **Renditions of this load characteristic seen on all major website systems.**
- **Servers are more likely to be throttled than networks.**

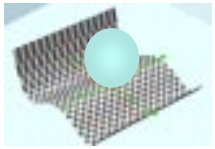


© 2000 Performance Dynamics Company™

Raw Data with Saturation



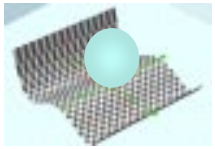
Fingers are "amputated" above a measured CPU utilization of 100%



© 2000 Performance Dynamics Company™

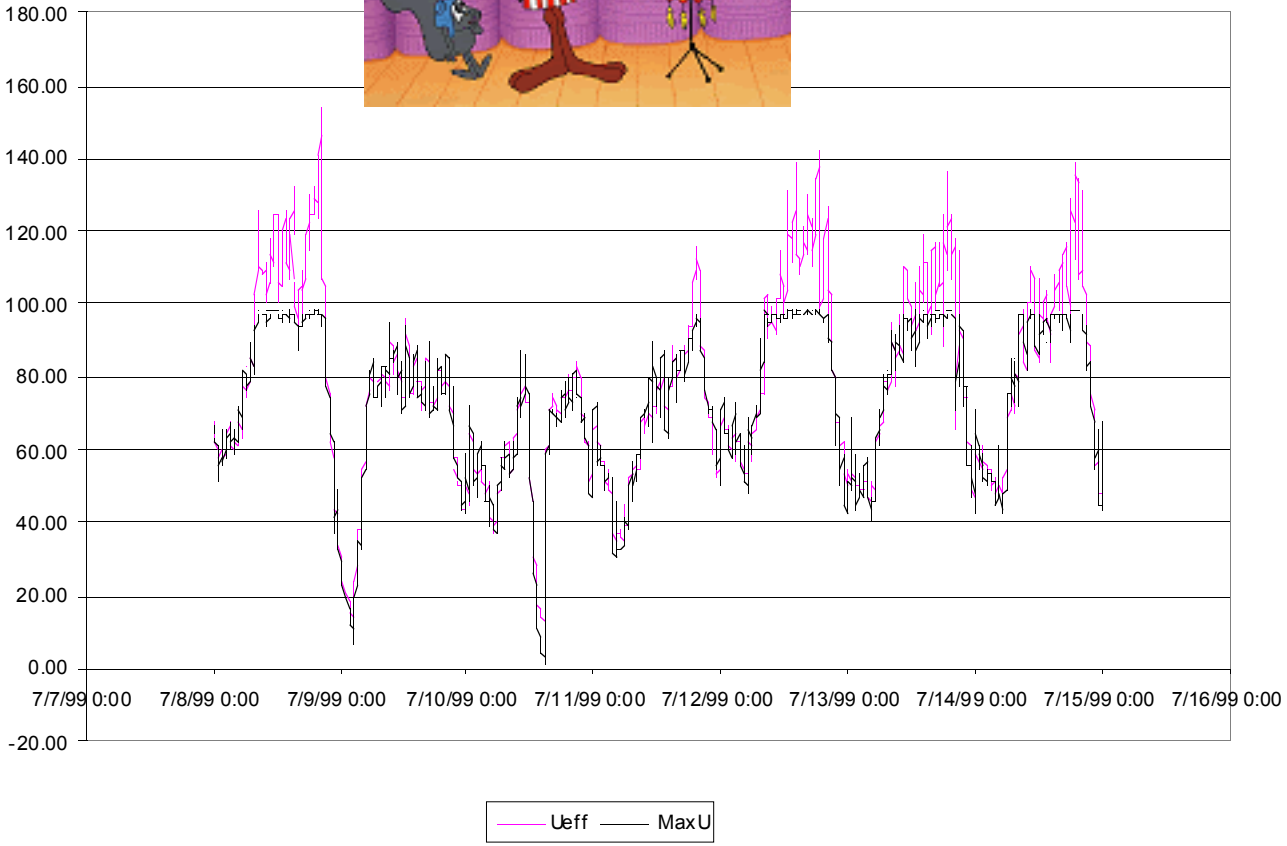
Effective Demand

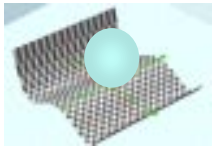
- **Would like to be able to predict what the "fingers" look like when they can't be seen in the CPU utilization data.**
- **We can do this with a multivariate statistical model.**
- **This is our first 'rabbit' !!**



© 2000 Performance Dynamics Company™

Predicted Signature

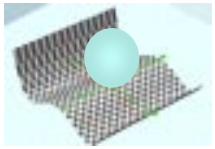




© 2000 Performance Dynamics Company™

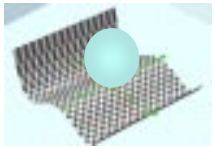
Capacity Metrics

- **Typical websites collect *cubic light-years* of performance data daily**
- **But what is the data trying to tell us?**
- **Raw data is like raw building material**
- **To build a building, you need a blueprint**
- **A performance *model* acts like a *blueprint***
- **Prototype capacity model built was in EXCEL/VBA**



© 2000 Performance Dynamics Company™

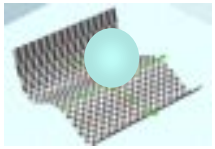
Nonlinear Regression



© 2000 Performance Dynamics Company™

E10K@333 MHz Headroom

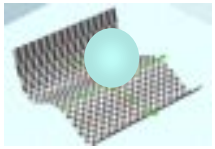
- **Current 52-way peaking around 95% utilized**
- **Peak to Avg. Ucpu ratio ~ 1.3**
- **Predicts that peak utilization will consistently hit 100% by end of Sept '99**
- **This is not necessarily bad**
- **But beware the SMP Wall!**



© 2000 Performance Dynamics Company™

Server Growth Model

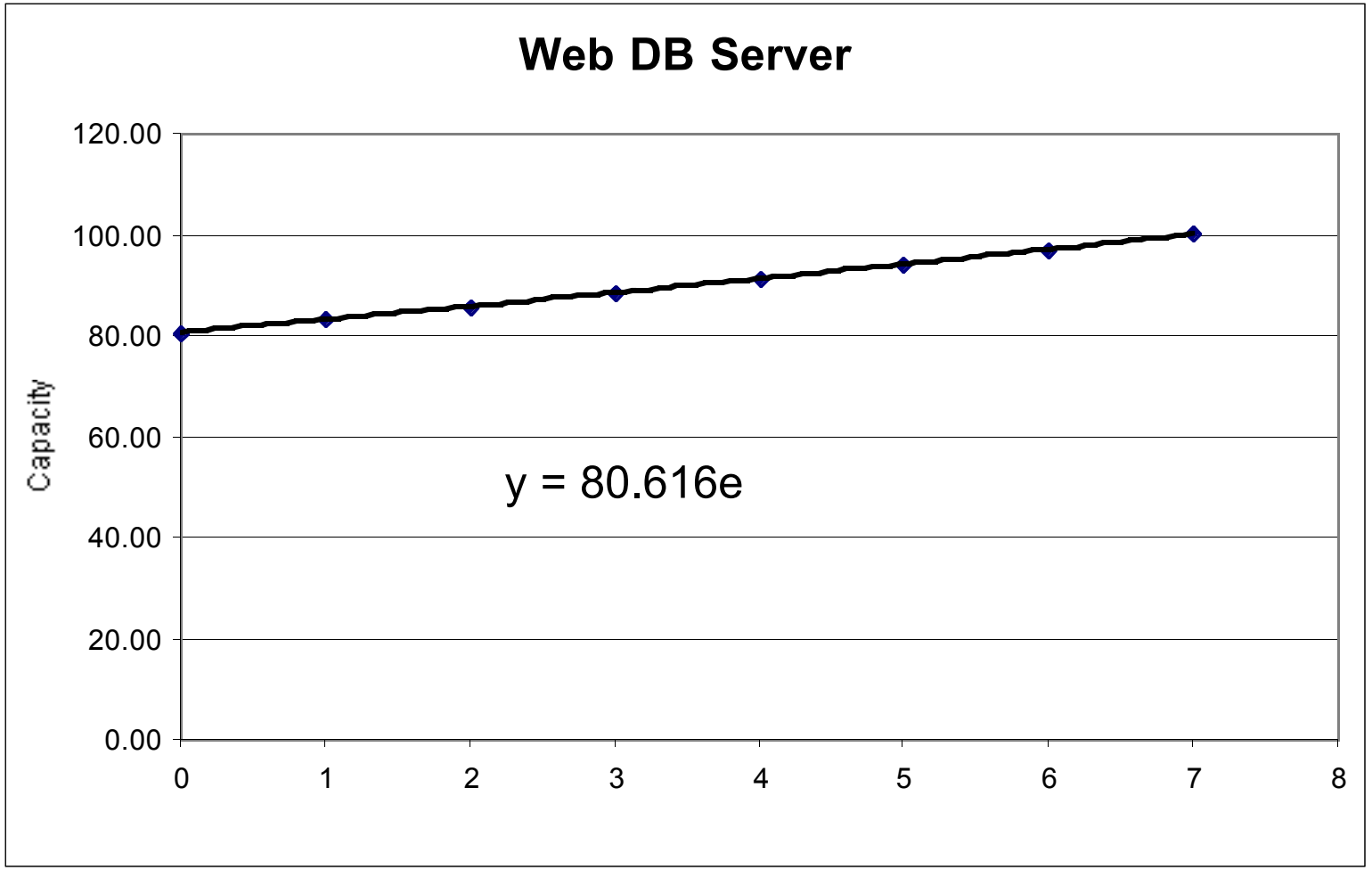
- **Expect server capacity consumption to exceed simple linear growth. As is evident in:**
  - **→ Business growth expectations (you want this!)**
  - **→ Network bandwidth measurements (MRTG)**
- **Compounding or *exponential* model**
- **Only a single parameter i.e., *growth rate***
- **Use EXCEL trending facility**

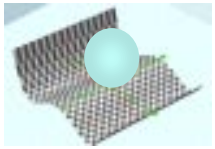


© 2000 Performance Dynamics Company™

Initial Fit to Data

$$C(W) = 80.62\, e^{0.0309W}$$

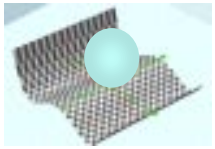




© 2000 Performance Dynamics Company™

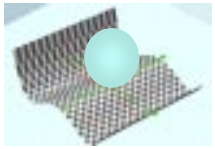
Growth Rate

- Fit to 7 weeks(W) of effective demand data
- Web DB server growth rate = 0.0309
- Doubles ea. Log(2)/0.0309 = 5.6 months !!
- 10 times faster than DP workloads
- 4 times faster than Moore's Law
- Agrees with other data at this website
- This is our second 'rabbit' !!!



© 2000 Performance Dynamics Company™

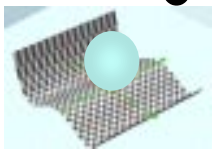
Server Bankrupcy



© 2000 Performance Dynamics Company™

More Realistic Scaling

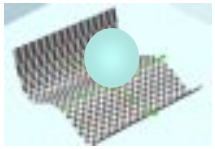
- Usesuper-serial scaling model (PPA Chap. 6)
- SS model has 2-parameters:
  - → Serial contention (σ) e.g., mutex locks
  - → Coherency(λ) e.g., cache fetch (rolloff factor)
- 64-way@400MHz M-values fit to SS model
  - → Contention: σ = 0.0061 (half a percent)
  - → Coherency: λ = 0.0016 (a fifth of 1 percent)
- 64-way@400MHz typical ORACLE data
  - → Contention: σ = 0.030 (about 3%)
  - → Coherency: λ = 0.002 (assume this doesn't change)
- Recalculate M-values using SS parameters
- Reduces headroom estimates
- Shortens the procurement mileposts



© 2000 Performance Dynamics Company™

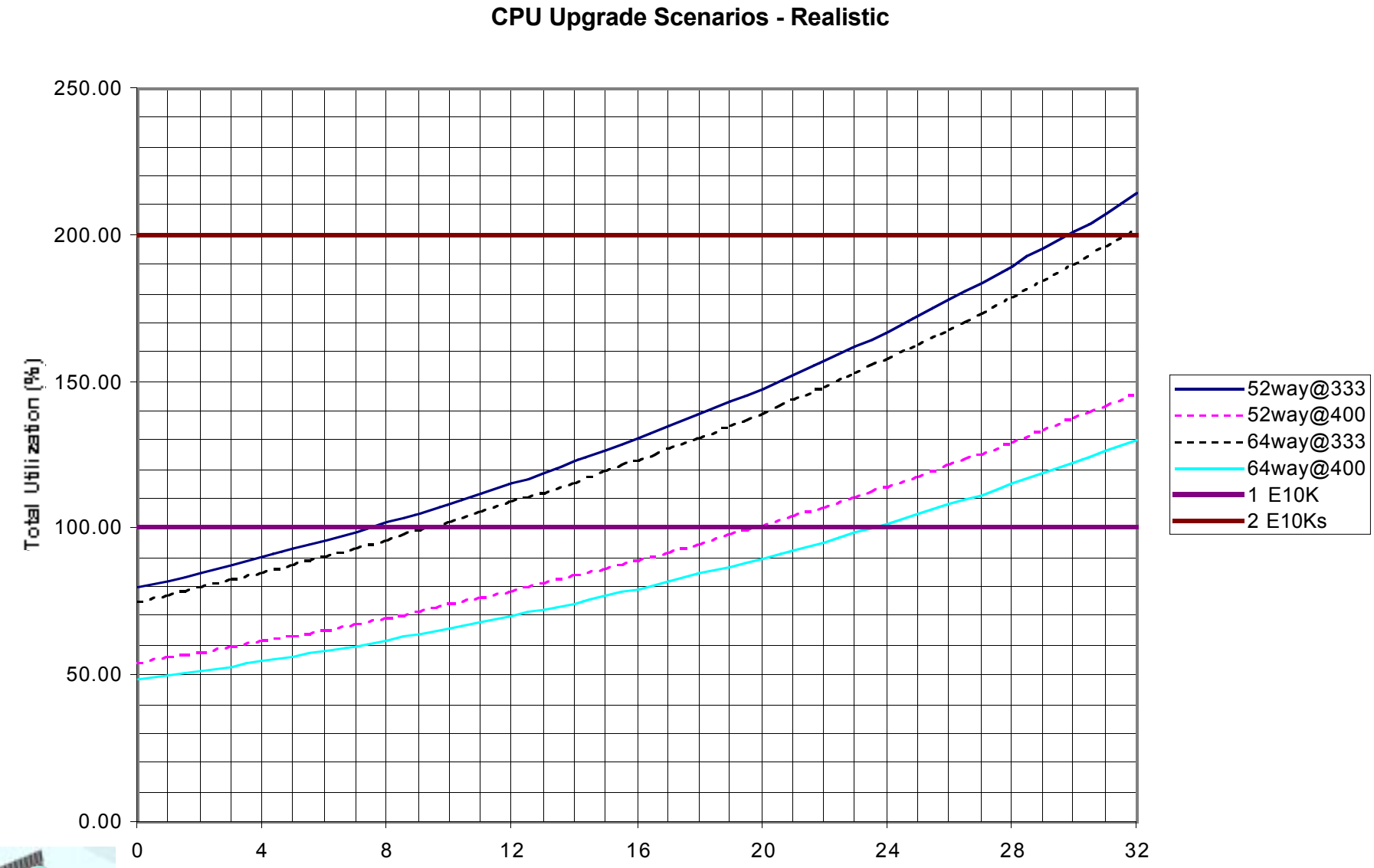
Server CPU Upgrades: II

- **Upgrade paths:**
  1. **Add more CPUs (52- to 64-way @ 333 MHz + 4MB cache)**
     - **Expect 6% headroom gain (cf. 15% previously)**
  2. **Faster CPUs (333 to 400 MHz) in a 52-way config**
     - **Expect 32% headroom gained (same)**
  3. **Do both: 64-way@400MHz + 8MB cache**
     - **Expect => 39% headroom gained (cf. 52% previously)**



© 2000 Performance Dynamics Company™

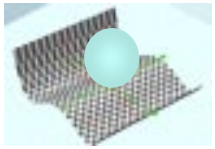
Realistic Growth



© 2000 Performance Dynamics Company™

Summary

- Use *multivariate regression* for short-term effects
  - → Daily data collection and *Effective Demand* model
  - → Weekly summarization
- Use *nonlinear regression* for long-term effects
  - → Used weekly summaries for *growth* model
- See Ch.6 in *The Practical Performance Analyst*, McGraw-Hill, 1998.
- If you'd like to learn more about this kind of capacity planning, consider attending Dr.Gunther's class *Guerilla Capacity Planning*, in August and November, 2000.
- Visit `http://www.perfdynamics.com` for more information about these and other topics.



© 2000 Performance Dynamics Company™