StackExchange

Open IT Operations and Stack Exchange's Environment

George Beech @GABeech
Kyle Brandt @KyleMBrandt
PICC 2011

Topics

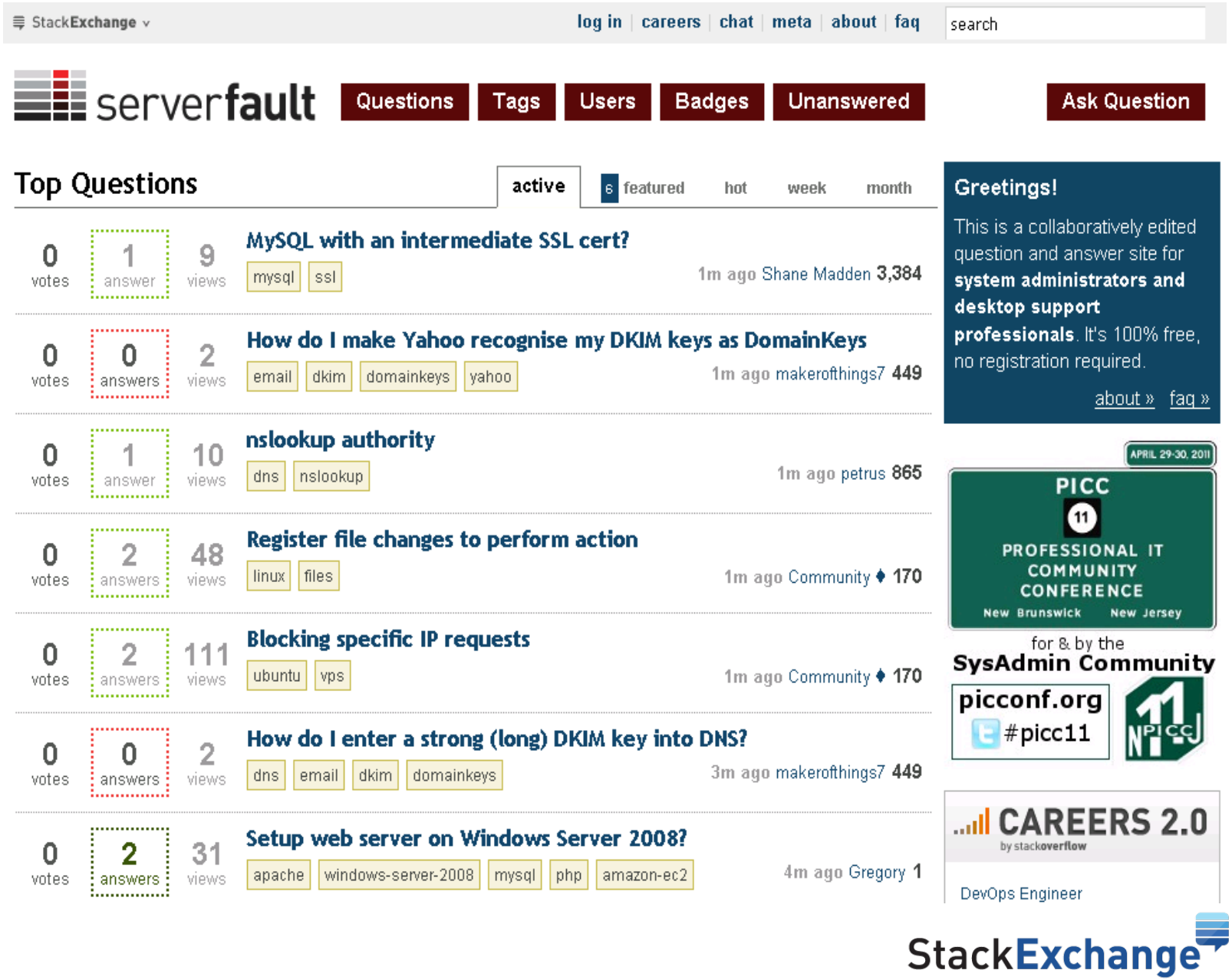
- Stack Exchange and our Philosophy of online community
- Our Infrastructure in a Nutshell
- Performance is Key
- Lessons learned

StackExchange

Stack Exchange

Stack Exchange is a growing network of 48 question and answer sites on expert topics from system administration to cooking to photography and gaming.

StackExchange


StackExchange
log in | careers | chat | meta | about | faq
search
serverfault
Questions
Tags
Users
Badges
Unanswered
Ask Question
Top Questions
active
featured
hot
week
month
0 votes
1 answer
9 views
MySQL with an intermediate SSL cert?
mysql
ssl
1m ago Shane Madden 3,384
0 votes
0 answers
2 views
How do I make Yahoo recognise my DKIM keys as DomainKeys
email
dkim
domainkeys
yahoo
1m ago makerofthings7 449
0 votes
1 answer
10 views
nslookup authority
dns
nslookup
1m ago petrus 865
0 votes
2 answers
48 views
Register file changes to perform action
linux
files
1m ago Community ♦ 170
0 votes
2 answers
111 views
Blocking specific IP requests
ubuntu
vps
1m ago Community ♦ 170
0 votes
0 answers
2 views
How do I enter a strong (long) DKIM key into DNS?
dns
email
dkim
domainkeys
3m ago makerofthings7 449
0 votes
2 answers
31 views
Setup web server on Windows Server 2008?
apache
windows-server-2008
mysql
php
amazon-ec2
4m ago Gregory 1
Greetings!
This is a collaboratively edited question and answer site for system administrators and desktop support professionals. It's 100% free, no registration required.
about »
faq »
APRIL 29-30, 2011
PICC
11
PROFESSIONAL IT COMMUNITY CONFERENCE
New Brunswick
New Jersey
for & by the
SysAdmin Community
picconf.org
#picc11
CAREERS 2.0
by stackoverflow
DevOps Engineer
StackExchange

Should network hardware be set to “autonegotiate” speeds or fixed speeds?

48

16

We recently had a little problem with networking where multiple servers would intermittently lose network connectivity in a fairly painful-to-resolve way (required hard reboot). This has been going on for about two weeks, seemingly at random, on different servers. No particular pattern that we could discern to it.

After some digging into it, we saw that the switch was reporting 100 Mbps for the problem port:

Auto	Disable	Disable	1000 Mbps Full Duplex	Link Up	Enable
Auto	Disable	Disable		Link Down	Enable
Auto	Disable	Disable	1000 Mbps Full Duplex	Link Up	Enable
Auto	Disable	Disable	1000 Mbps Full Duplex	Link Up	Enable
Auto	Disable	Disable		Link Down	Enable
Auto	Disable	Disable		Link Down	Enable
Auto	Disable	Disable	1000 Mbps Full Duplex	Link Up	Enable
Auto	Disable	Disable	1000 Mbps Full Duplex	Link Up	Enable
Auto	Disable	Disable	100 Mbps Full Duplex	Link Up	Enable
Auto	Disable	Disable	1000 Mbps Full Duplex	Link Up	Enable
Auto	Disable	Disable	1000 Mbps Full Duplex	Link Up	Enable

This sounds remarkably like what happened in the Joel Spolsky article Five Whys

> Michael spent some time doing a post-mortem, and discovered that the problem was a simple configuration problem on the switch. There are several possible speeds that a switch can use to communicate (10, 100, or 1000 megabits/second). You can either set the speed manually, or you can let the switch automatically negotiate the highest speed that both sides can work with. **The switch that failed had been set to autonegotiate. This usually works, but not always, and on the morning of January 10th, it didn't.**

We have now **disabled auto-negotiate** on our network hardware and set it to a fixed rate of 1000 Mbps (gigabit).

My questions to those with more server hardware networking expertise:

1. How common are auto-negotiate problems with modern networking hardware?
2. Is it considered good, standard networking practice to disable auto-negotiate and set fixed speeds when setting up networking?

tagged

networking × 3175

ethernet × 164

asked
1 year ago

viewed
4,436 times

latest activity
9 months ago

18 Answers

active | oldest | **votes**

78

1. I have yet to see a problem with auto-negotiation of network speeds that isn't caused by either (a) a mismatch of manual on one end of the link and auto on the other or (b) a failing component of the link (cable, port, etc).
2. This depends on the admin, but my experience has shown me that if you manually specify the link speeds and duplex settings, than you are bound to run into speed mismatches. Why? Because it is nearly impossible to document the various connections between switches and servers and then follow that documentation when making changes. Most failures I have seen are because of 1(a) and you only get in to that situation when you start manually setting speed/duplex settings.

As mention in the Cisco documentation:

> If you disable autonegotiation, it hides link drops and other physical layer problems. Only disable autonegotiation to end-devices, such as older Gigabit NICs that do not support Gigabit autonegotiation. Do not disable autonegotiation between switches unless absolutely required, as physical layer problems can go undetected and result in spanning tree loops.

Unless you are prepared to setup a change management system for network changes that requires the verification of speed/duplex (and don't forget flow control) or are willing to deal with occasional mismatches that come from manually specifying these settings on all network devices, then stick with the default configuration of auto/auto.

In the future, consider monitoring the errors on the switch ports with MRTG so you can spot these issues before you have a problem.

Edit: I do see a lot of people referencing negotiation failures on old equipment. Yes this was an issue a long time ago when the standards were being created and not all devices followed them. Are your NICs and switches less than 10 years old? If so, then this won't be an issue.

mod | link | edit | delete | flag

edited **Jan 25 '10 at 19:21**

answered **Jan 25 '10 at 19:15**

Doug Luxem

5,119 ●1 ●11 ●40

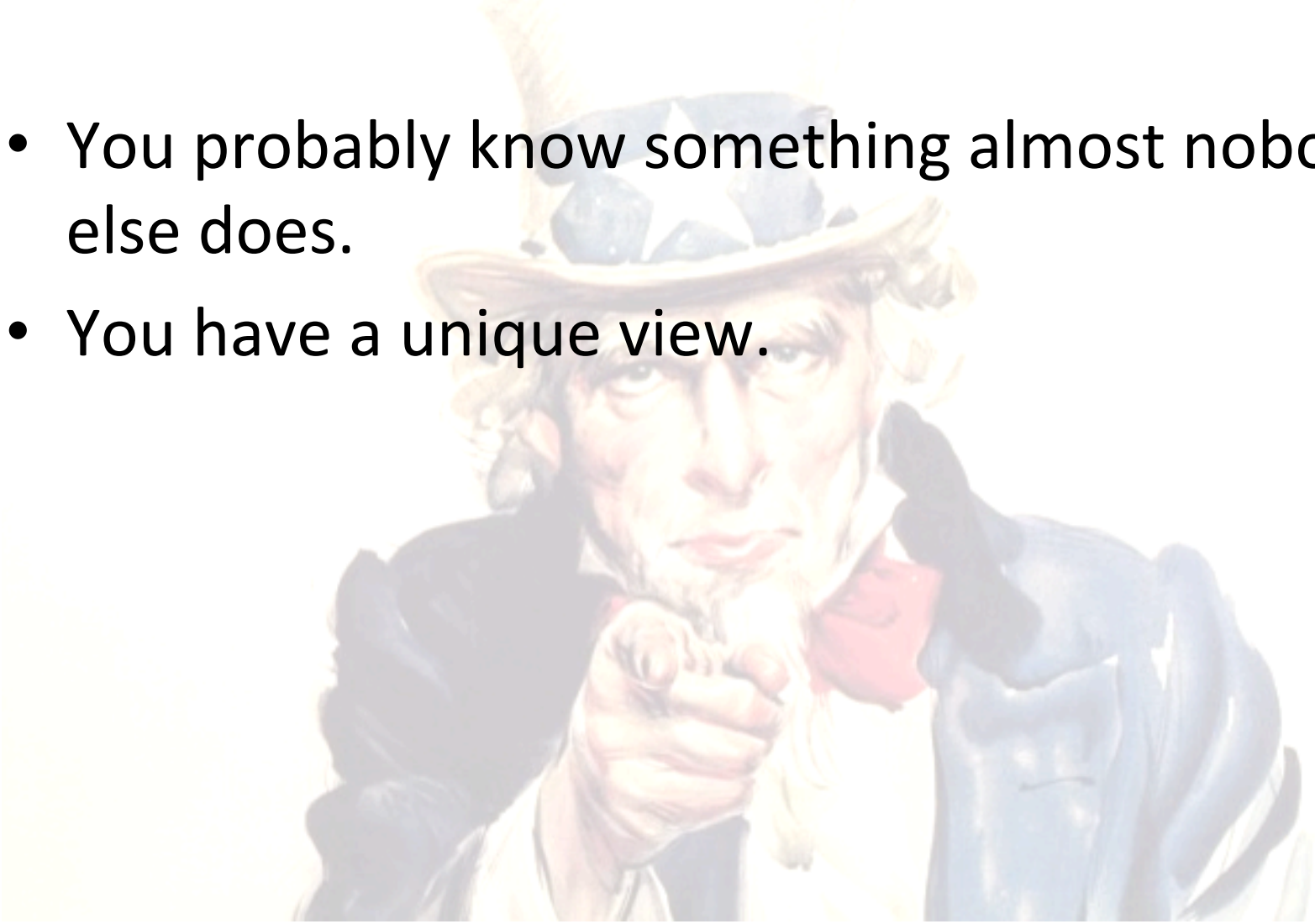
Why Participate online

- The System Administration Community Needs you
- Its good for you

StackExchange

We Want You

- You probably know something almost nobody else does.
- You have a unique view.



StackExchange

It is good for you

- More fluency and facility
- Interview Skills
- You will become a better writer

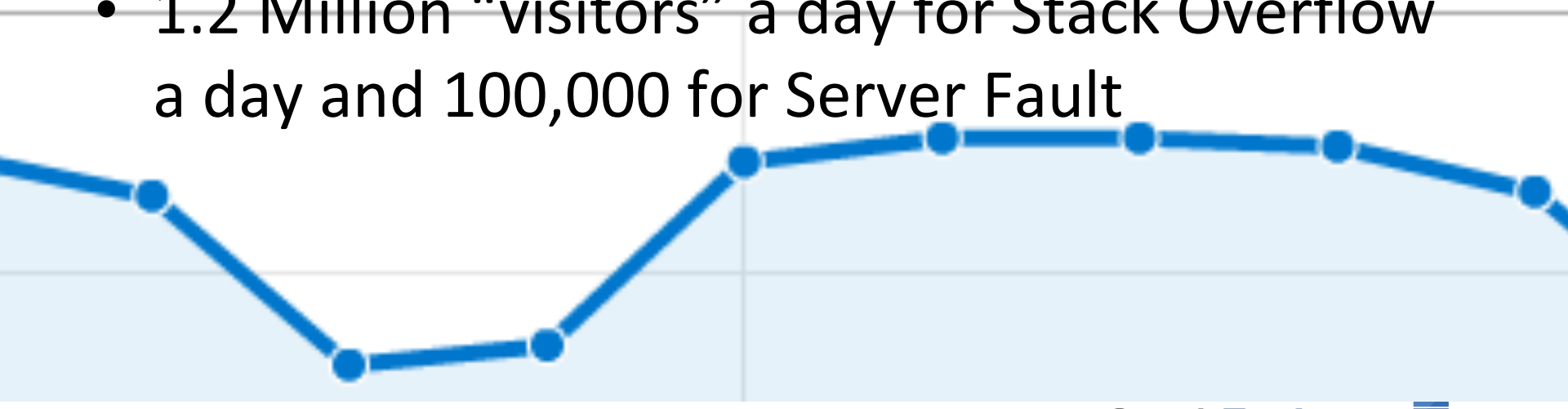
StackExchange

Why have Open IT Operations?

- Betters decisions
- Helps your field
- Security by Obscurity

Stack Exchange Stats

- 120 million page views a month
- Network number 250 in the US
- 800 HTTP requests a second
- 180 DNS requests a second
- 1.2 Million “visitors” a day for Stack Overflow a day and 100,000 for Server Fault

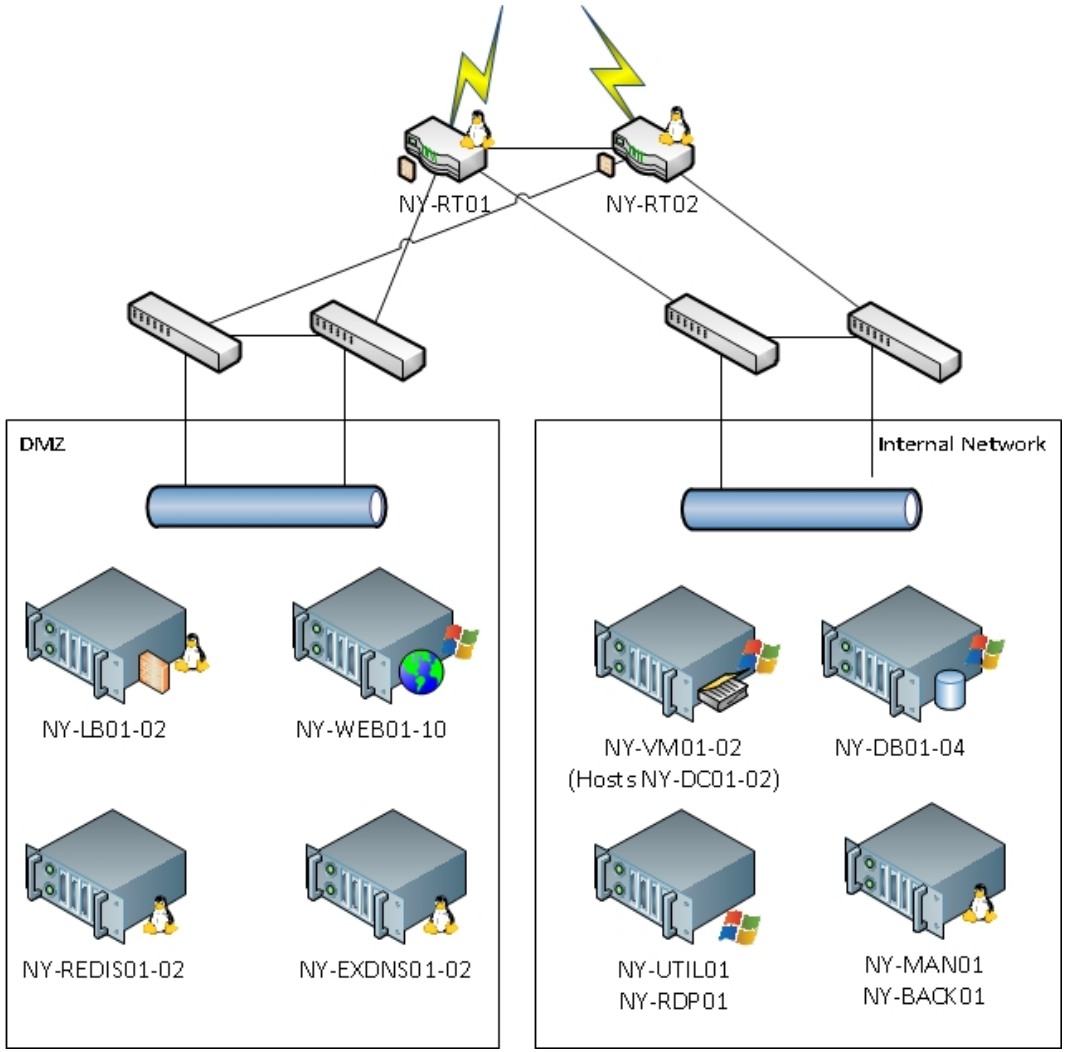


StackExchange

What is Stack Exchange's Core Built On?

- Largely a Microsoft stack using .NET MVC Razor, IIS, and SQL Server
- Linux HAProxy and Redis
- Awesome Programmers

Network Diagram



StackExchange

This is a transition

And it goes on and on my friend

StackExchange

Performance Is A feature

"it is well known that speed correlates with activity, the faster you are, the more there is, and the SLOWER you are, the less there is … bottom line, performance is a feature. And a pretty important one."

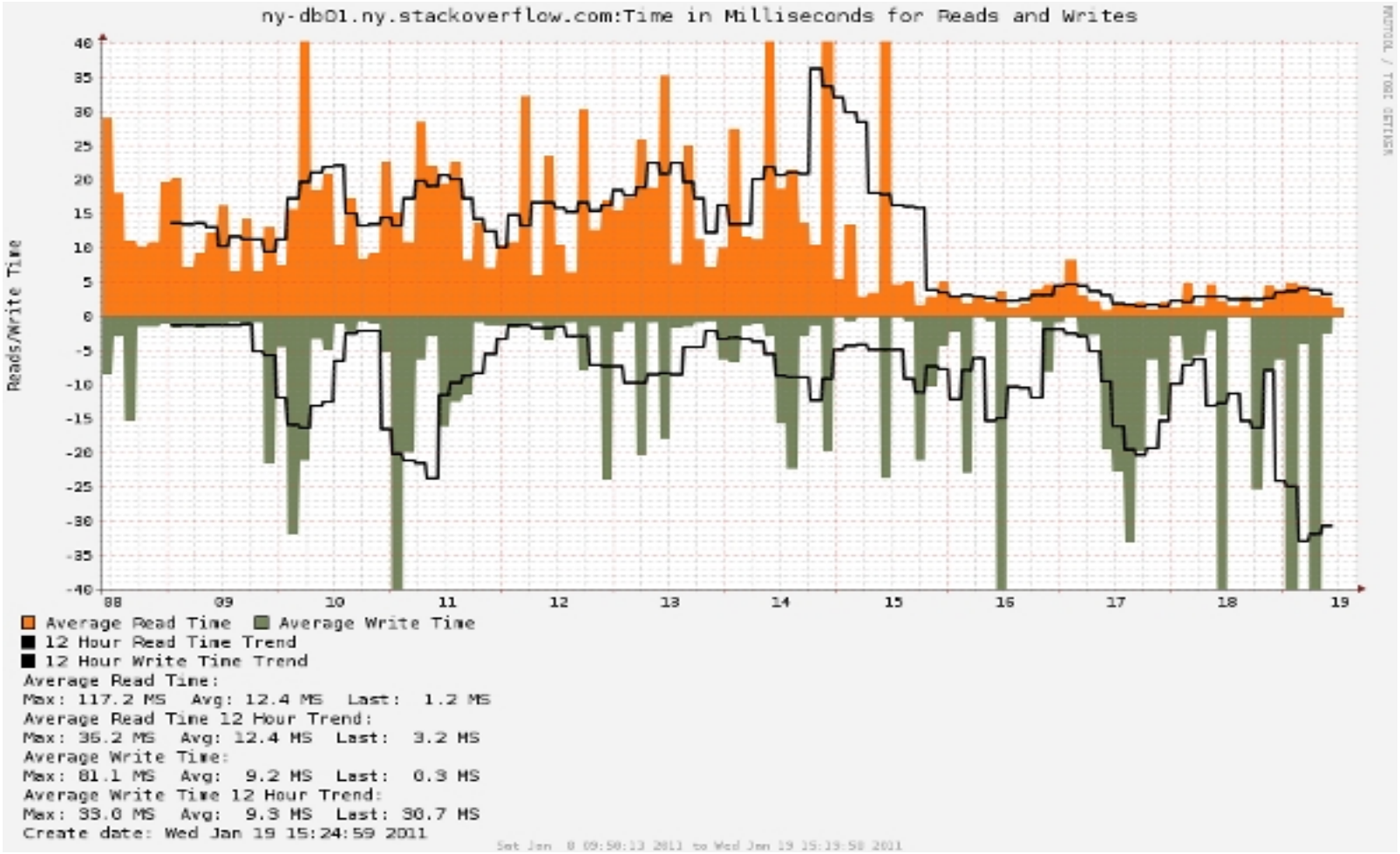
- Jeff Atwood

StackExchange

Common Bottlenecks

- Disk
- CPU
- Network

NOT good performance



StackExchange

Disk Performance

- For DB servers, this is key
- Evaluated Options
  - SAN
  - Disk Enclosure
  - SSD
  - SSD on PCI (i.e. FusionIO)

DAS Enclosure

- Drive Enclosure, Directly Connected
- Pro
  - Large Number of Spindles
  - Relatively Low cost
- Con
  - Limited Flexibility
  - Still .. Kind of expensive for what you get

SANs

- Pro
  - Very Flexible
  - Generally Expandable
- Con
  - BOHICA Expensive
  - If you don't have the infrastructure, you need to build it
  - Highly Specialized Configuration

PCI Flash Drives

- Fusion IO Type Drives
- Pro
  - Oh my, that's fast
  - Price tolerable
- Con
  - New Tech
  - No good SPoF Protection

StackExchange

SSDs

- Normal, SATA SSDs, well almost
- Pro
  - Fast, we are talking flash here
  - Flexible
  - “Cheap”
- Con
  - If you by from your vendor, it’s not worth it
  - If you don’t buy from your vendor, they aren’t under warranty.

SSD vs Fusion IO

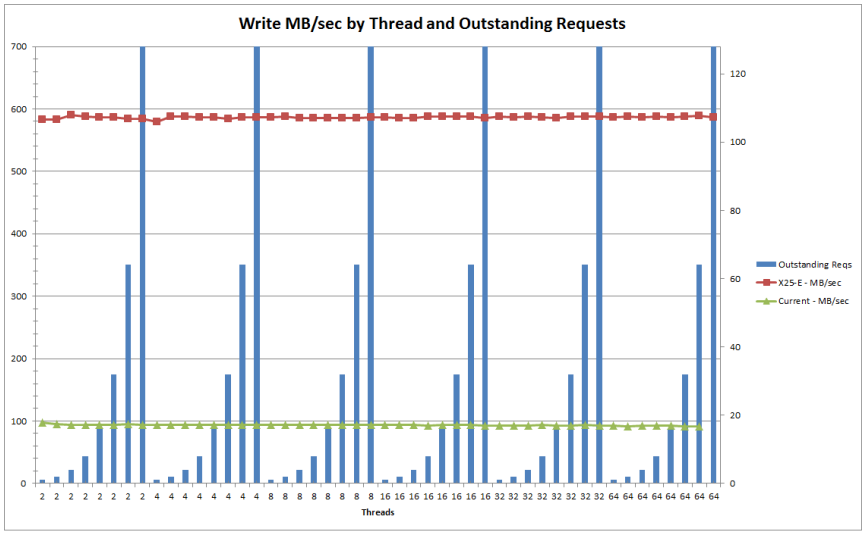
Random Reads — 2 threads, 8 outstanding requests, 64k blocks

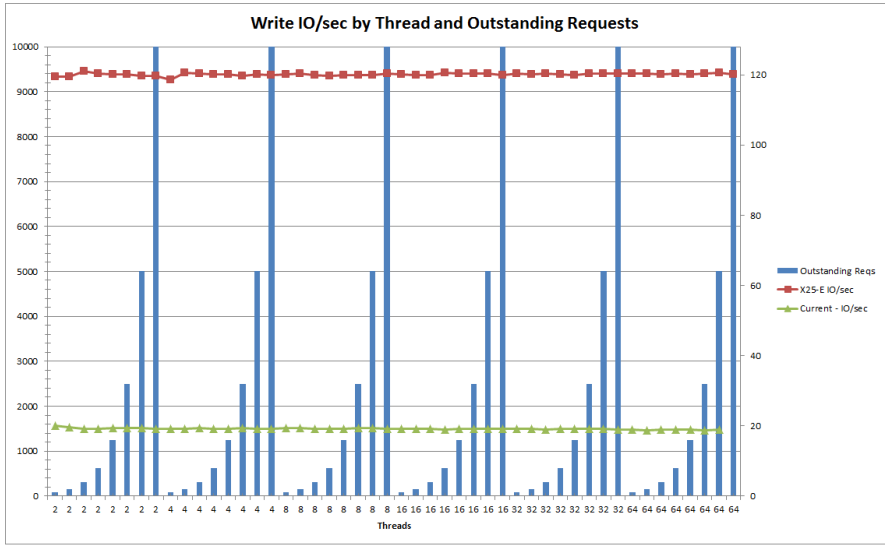
	Fusion IO	Intel X25-E
MB/s	1424	1064
IOPS	22788	17023

Random Writes — 2 threads, 1 outstanding request, 64k blocks

	Fusion IO	Intel X25-E
MB/s	632	584
IOPS	10114	9337

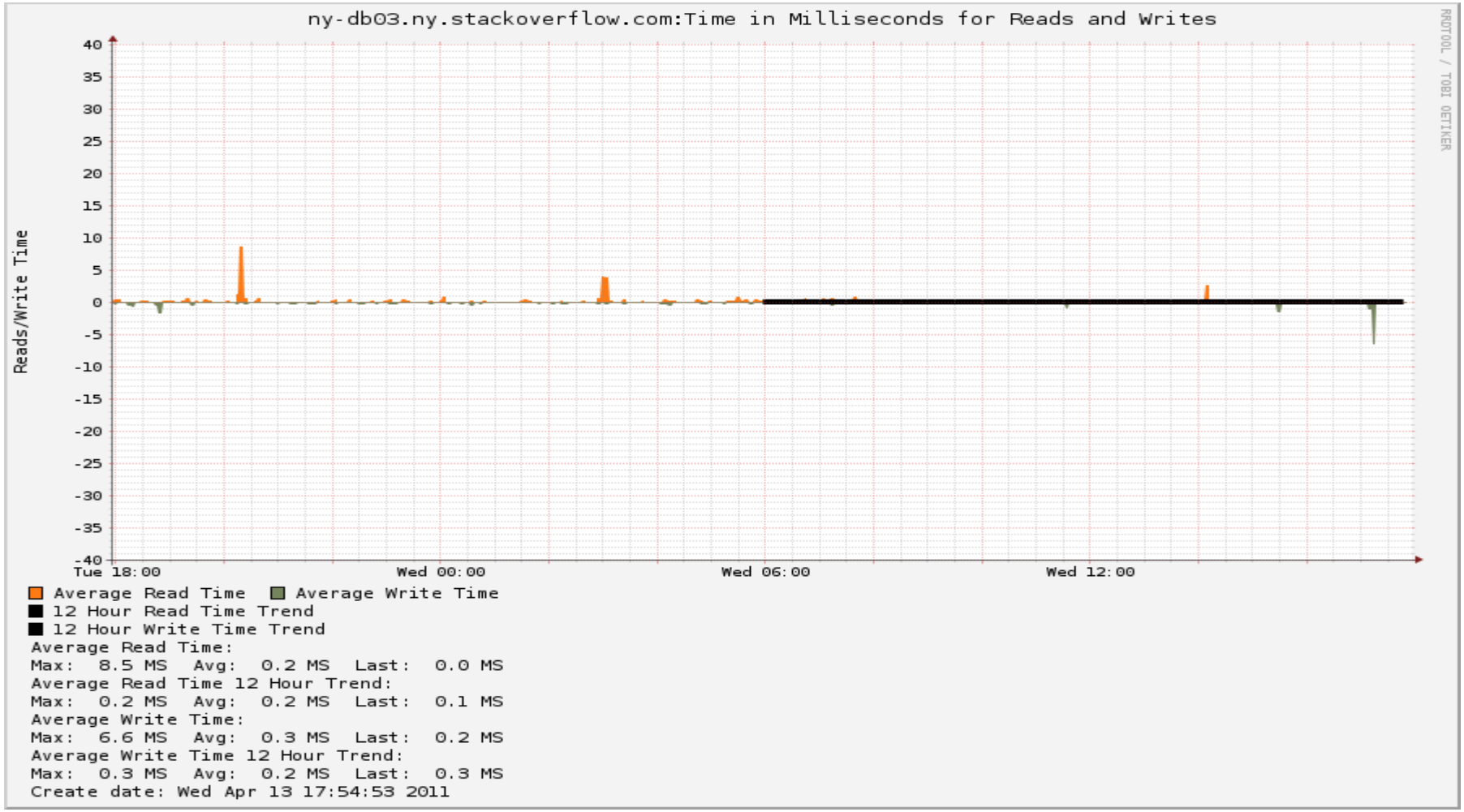
SSDs Win





StackExchange

GOOD performance



StackExchange

SSDs Everywhere

- Vendor Prices Suck
- We have decided that taking on the risk of non-warranty covered disks is ok
- Everywhere we can get some performance out of a better disk system, we will put in SSDs
- Intel rocks the house
- The new 3rd Gen technology from Intel gives you more storage, and better performance in the MLC format

StackExchange

Network Performance

- Weird Network Behavior
- LOTS of 0 length TCP windows
- Random failures
- Could not instrument our switches
- If your network is slow, it doesn't matter how fast your machines are

You get what you pay for

- Pay the name brand premium
- We chose cisco because we know the equipment and IOS
- Dell switches are cheap, but you get cheap equipment
  - No instrumentation
  - Not true wire-speed gig-E (on all ports)
  - NO INSTRUMENTATION

Intel I/OAT

- Intel® QuickData Technology — enables data copy by the chipset instead of the CPU, to move data more efficiently through the server and provide fast, scalable, and reliable throughput.
- Direct Cache Access (DCA) — allows a capable I/O device, such as a network controller, to place data directly into CPU cache, reducing cache misses and improving application response times.

Intel I/OAT (cont)

- Extended Message Signaled Interrupts (MSI-X) – distributes I/O interrupts to multiple CPUs and cores, for higher efficiency, better CPU utilization, and higher application performance.
- Receive Side Coalescing (RSC) — aggregates packets from the same TCP/IP flow into one larger packet, reducing per-packet processing costs for faster TCP/IP processing.
- Low Latency Interrupts — tune interrupt interval times depending on the latency sensitivity of the data, using criteria such as port number or packet size, for higher processing efficiency.

Lessons Learned

StackExchange

Oops, Naming is Hard

- I picked the wrong Active Directory Name Twice:
  - stackoverflow.com – Don't use an actual domain
  - ny.stackoverflow.com – To Concrete

StackExchange

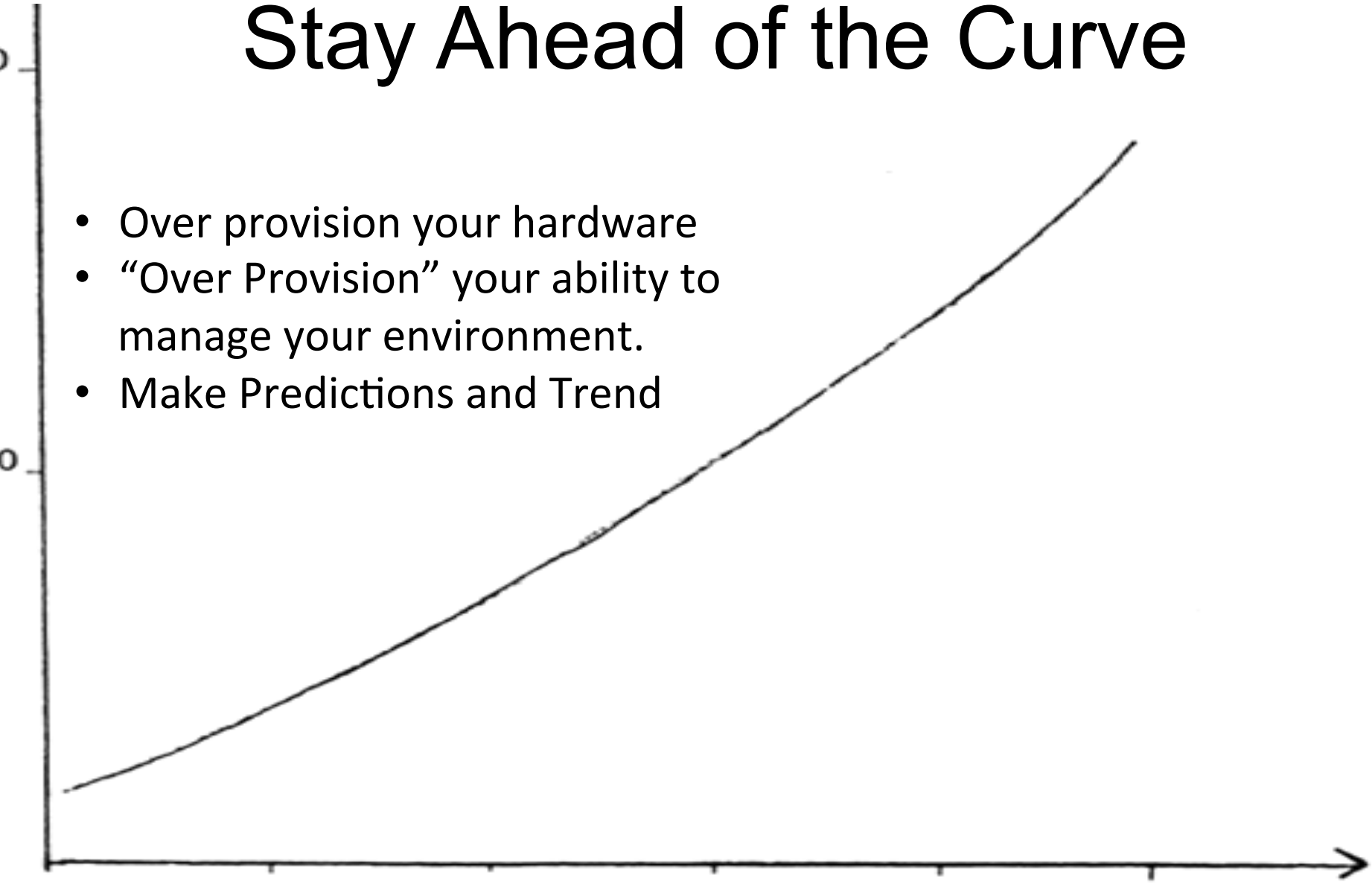
Don’t Forget about Power

- Overload on Failure
- Solution, 2/3rds capacity at power loss for web servers:
  - Web01: Two Power Supplies in both A and B feeds
  - Web02: Feed A only
  - Web03: Feed B only
  - etc…

StackExchange

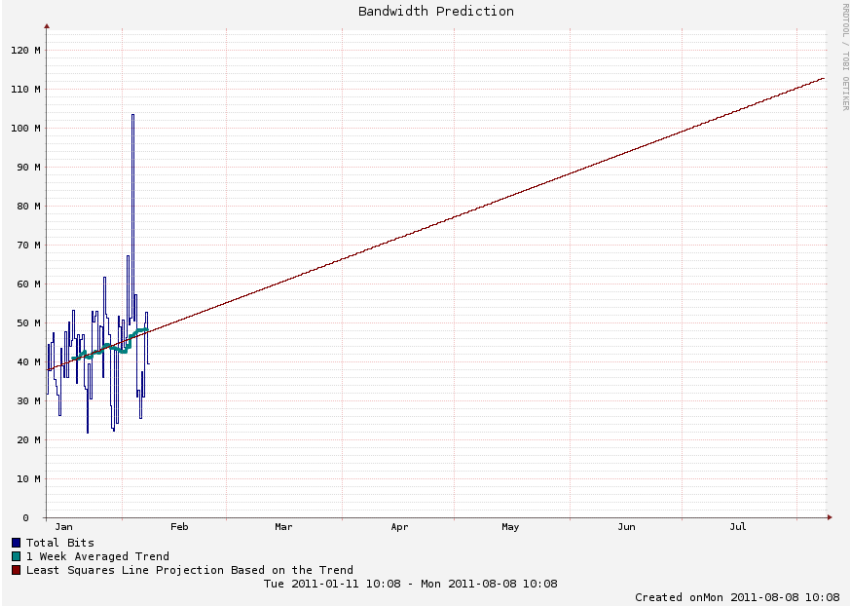
Stay Ahead of the Curve

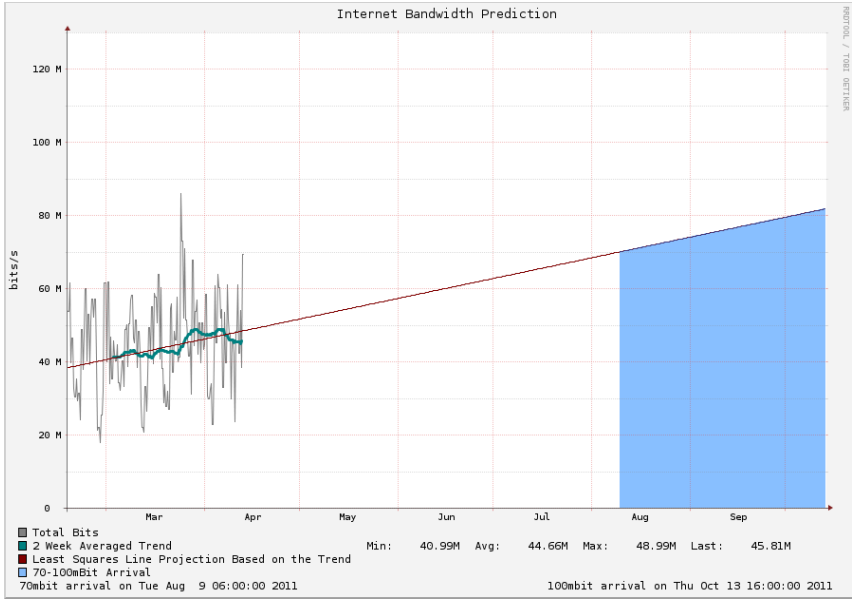
- Over provision your hardware
- “Over Provision” your ability to manage your environment.
- Make Predictions and Trend



StackExchange

Bandwidth Predictions





StackExchange

Don’t Save

- Starting small will end up costing you.

StackExchange

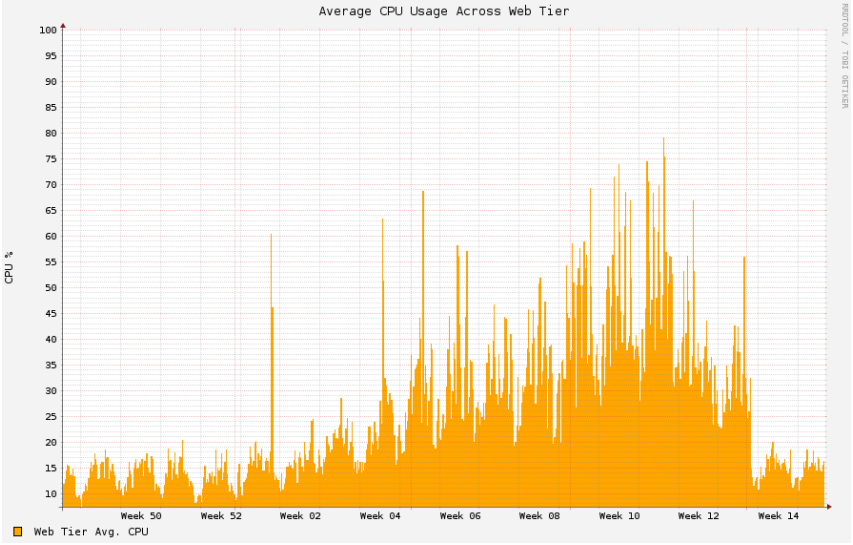
More Data is More Awesome

- Collect what you can as soon as possible
- When you have data, you can stop guessing
- Know what you are looking at., i.e. Max vs Avg

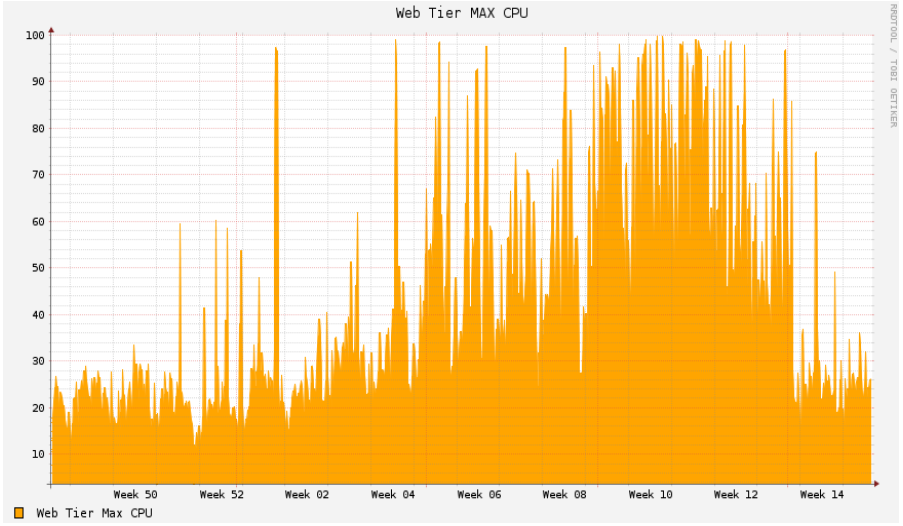
StackExchange

Max Vs. Average

Average CPU: 5 minutes samples of 1 minute average of All servers from web tier



Max CPU: MAX of all 1 minute averages from all servers in the web tier. This reflected when we actually started to see issues



Coding is Not Optional

- Communicate better with your developers
- There isn't always the tools that you need
- It will hold you back

QUESTIONS?

AMA

StackExchange

Additional information

http://serverfault.com
http://stackexchange.com/about
http://blog.serverfault.com
http://www.intel.com/go/ioat

StackExchange