



Application Front-Ends: The Key to Scaling Data Centers?

Steve Shah

Director of Product Management, Security Products
Citrix Systems, Inc.



Questions to Answer



1. How do I turn my strategic needs into technological choices?
2. How do the various AFE technologies impact the business?
3. What is the most effective way of measuring AFEs?

4. Is Citrix NetScaler the greatest AFE ever?



Resources in the Data Center



What Moves the Numbers?

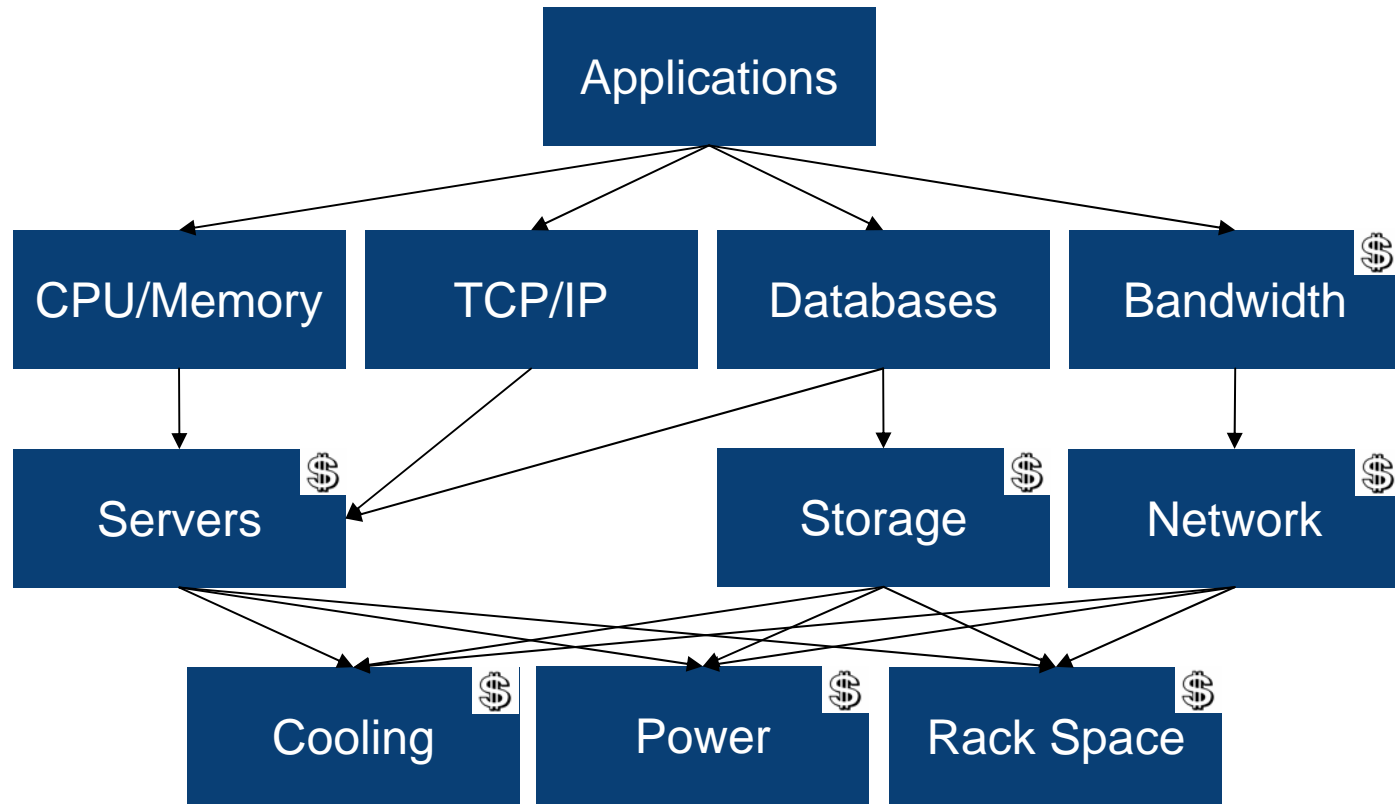
Revenue

Expenses

- Advertisements?
- Product Sales?
- Service Sales?
- Referrals?

- Data center costs?
- Infrastructure costs?
- Help desk calls?
- Visits to the branch?

Datacenter Costs



Will the items that translate into money spent speak up?

Business Needs Into Tech Needs



What technological metrics impact the revenue and expenses?

If AFEs aren't impacting either revenue or expenses, they're just a big, pretty toys.

Faster response times?

Reducing bandwidth needs?

Support

(Although as a vendor, I support the purchase of big, pretty toys.)

Purchases?

Enabling more interactive applications?

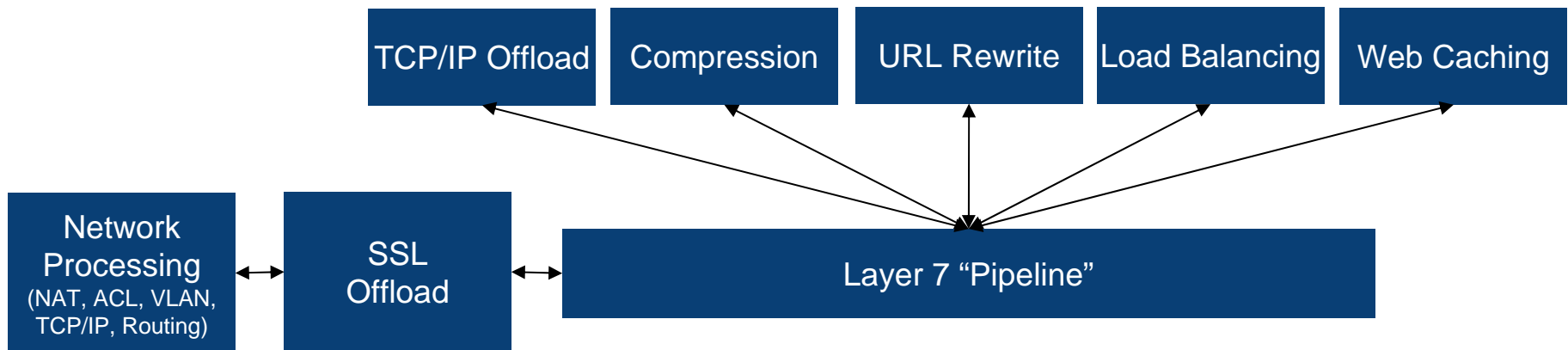
Reducing storage requirements?



Breaking Down the AFE and its Benefits

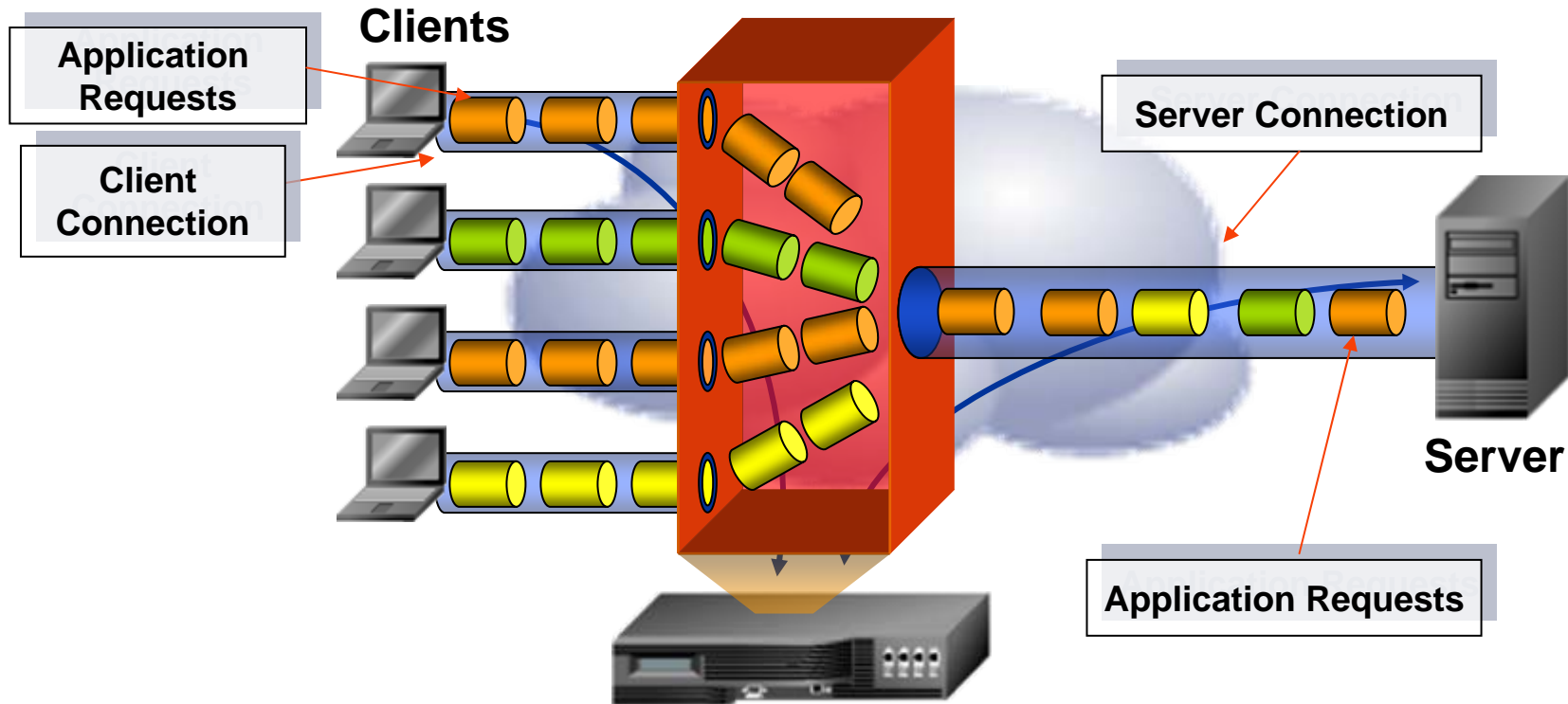


Typical AFE Architecture



- Integrated AFEs don't repeat work
 - Saves on latency as traffic bounces from box to box
 - No double processing of SSL, TCP/IP, or HTTP
- Incremental cost/load of a feature is much lower
 - Adding web caching is cheap if you've already processed HTTP for SLB
 - Ongoing management cost

TCP/IP Offload



- Allows multiple HTTP requests on one TCP connection
- Compensates for HTTP's use of many short lived connections
- Long lived connections perform better and reduce server load

Impact of TCP/IP Offload



- Long lived connections perform better
 - Eliminates costly connection setup/teardown
 - Allows TCP window size to grow to maximum size
 - Van Jacobson gives a fast path to the application
 - Small, static objects benefit the most, large dynamic responses the least
- Lower TCP/IP overhead means faster servers
 - Reduced CPU load means faster response time
 - Reduced CPU load means more CPU available for applications
- Faster servers means fewer are necessary
 - Option to either remove servers or serve faster
 - Determine what is desired more: additional revenue or lowered cost

Compression

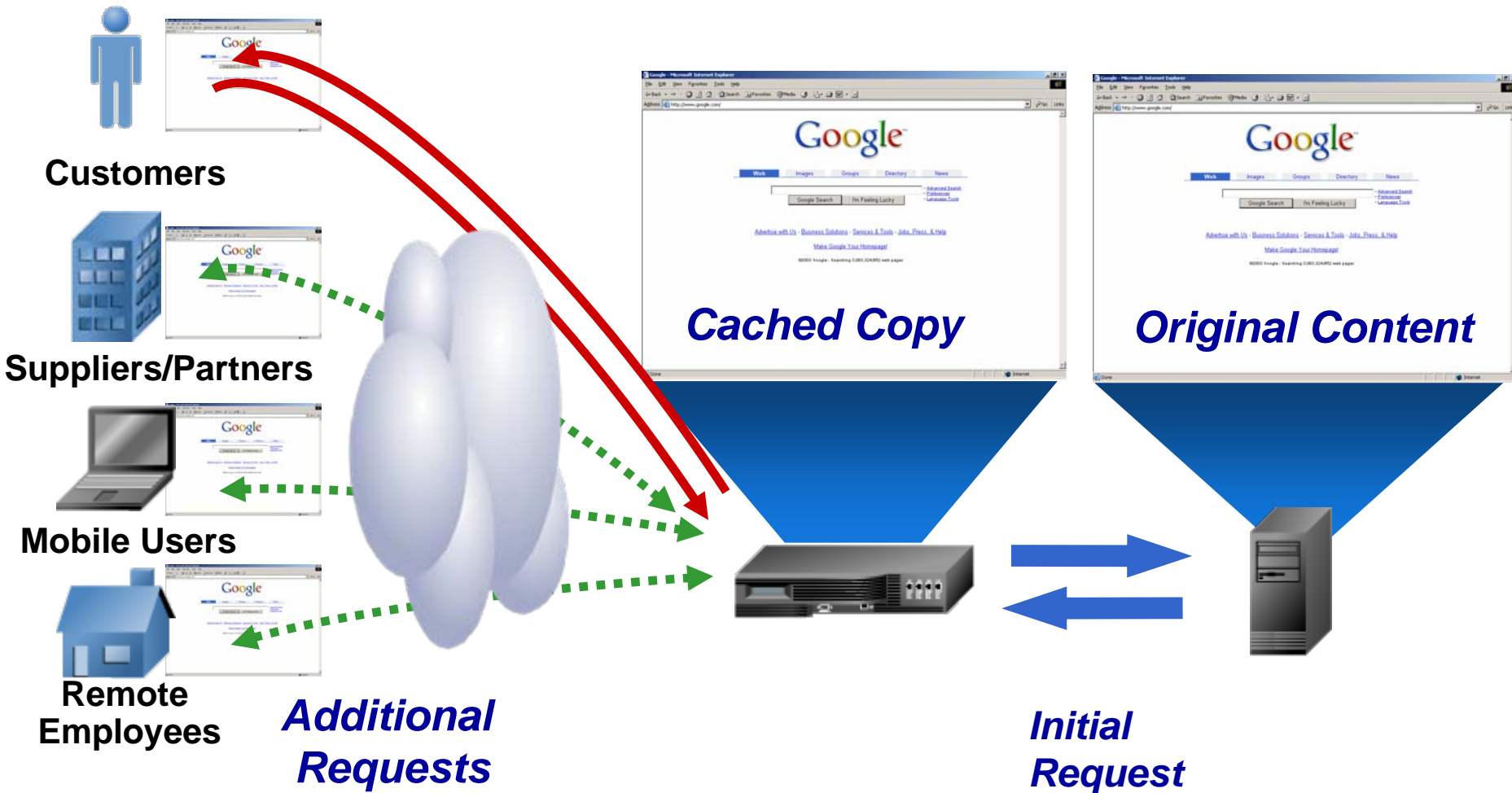


- Compression reduces bandwidth consumption
 - Clientless compression is up to 5x with Gzip and 15x for differential
 - Depending on mix of images, compression averages to 30-70%
- Compression also provides faster response times
 - Higher latency and low bandwidth users feel the greatest impact
 - End user perception can be significant
- Compression on AFEs can offload server load
 - Impact to server CPU can be upwards of 20-30%

- **SSL Acceleration significantly reduces server load**
 - Over 50% CPU savings are not uncommon
 - Compliance requirements for SSL to the server still benefits 20-30%
- **SSL Acceleration improves response times**
 - Accelerated RSA math results in faster initial response
 - TCP/IP offload, compression, and layer 7 SLB can now occur
- **Reduced cost of administration**
 - One point for SSL certificate administration
 - One point for security patching/updates
 - Far more effective use of SSL acceleration hardware

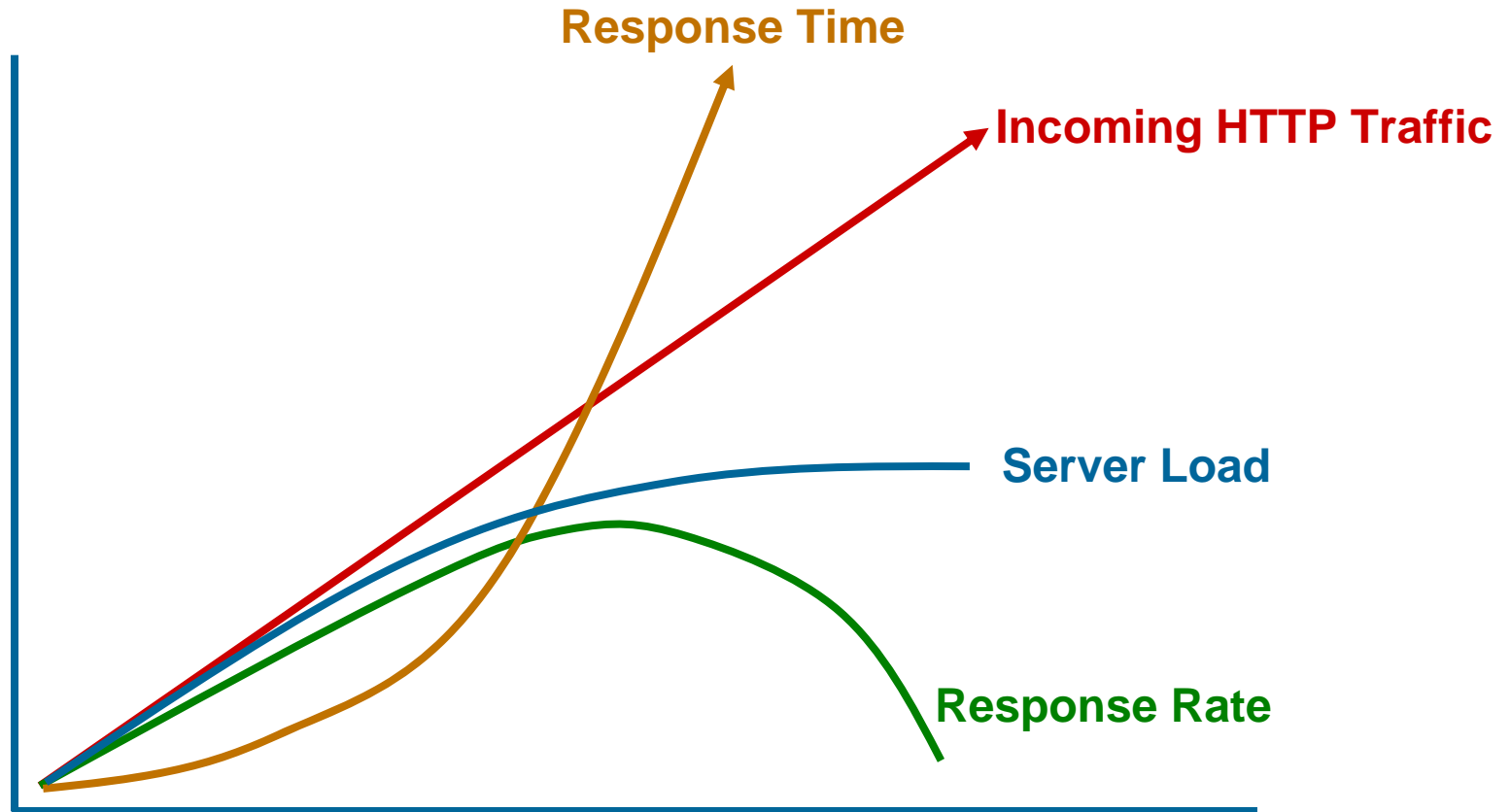


Web Caching

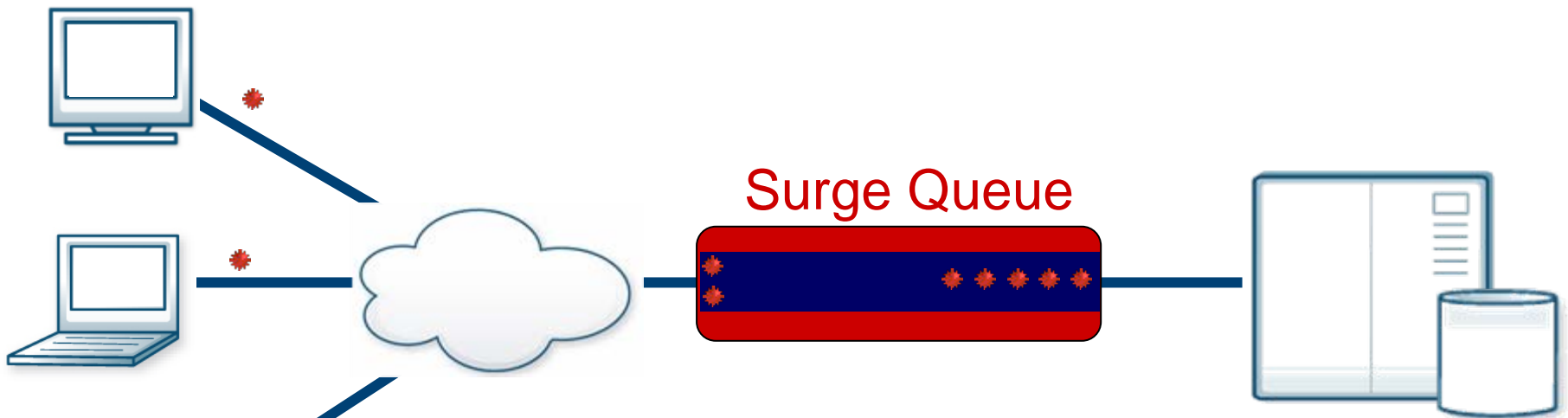


- Cache hits can completely eliminate server load
 - Most static content can be cached for long periods of time
- “But my site is all dynamic...”
 - Unlikely. Every .css, .js, .gif, .jpg, etc. can be cached
 - Dynamic caching can further improve hit rates
 - Reports can be cached
 - Pages that don’t change for a few minutes (e.g., sports scores) can be cached
 - Search results can be cached
 - Caching dynamic data can significantly reduce database load
- RAM based caches can reduce response times
 - Very fast response times
 - Even smaller sizes provide very high hit rates (90/10 rule)

Surge Protection: The Problem

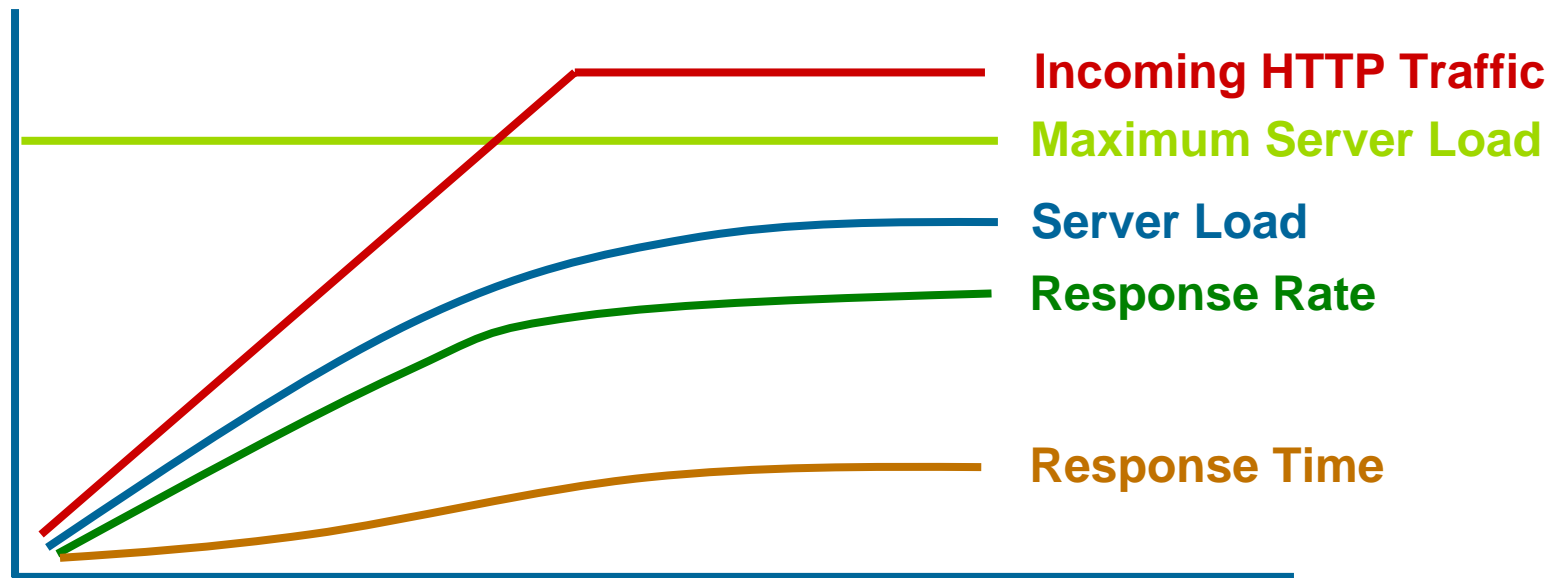


What Surge Protection Does



- Maintains optimal traffic rate to the server
- Does not drop valid traffic surges
- Results in optimal end user response time

Surge Protection: Result



- End to server spirals
 - Costly outages caused by cascading server failures stopped
 - Existing servers are effectively used – no CPU wasted on juggling
- Performance improvements
 - End user experience improves, especially during significant events
 - Consistent, predictable behavior improves site stickyness



Tips from Experience



Running a Smooth Test



- Enumerate the data you want to collect
 - Insures you're getting all the data you need
 - Provides consistency in analysis between vendors
- Make sure your questions can be quantified
 - This gives a much simpler way of comparing features
 - Numbers don't "spin"
- If possible, collect multiple samples
 - Error bars give you a sense of how consistent the performance is
 - Error bars let you identify "noise" and account for it
 - Be sure to reboot between samples to get consistency

- Test with a full configuration
 - Configuration file sizes can impact performance
 - ACLs, VLANs, routes, policies, etc. can have an impact on performance
- Test in similar network connections as real life
 - How many concurrent connections do you expect?
 - How many client IPs do you expect?
 - What kind of port scanning/DoS traffic do you expect?
 - What mix of object sizes do you expect?
- Watch for saturated test tools
 - Has the load generator peaked? Web server?
 - Are the switches dropping packets?
 - Is Link Aggregation distributing traffic evenly?

Big Throughput vs. Latency



- Two ways to look at improvements by AFEs
 - What increase in throughput is possible?
 - What reduction in latency for a response is possible?
- Increase in throughput
 - Requires the use of significant load generation (Spirent, IXIA, etc.)
 - curl, ab, http_load and friends don't generate enough traffic
- Reduction of latency
 - Doesn't require significant traffic to test
 - Latency (with and without) can be determined with a packet capture



Case Studies





Challenge

- Scale Outlook Web Access to 10,000 users

Solution

- SSL acceleration, layer 7 switching, TCP/IP offload, HTTP compression, and DDoS protection

Results:

- Reduced server count by 80%
- Significant performance increase
- Total cost savings of >\$200k
- ROI in less than 3 months



Document
Management
System



Challenge

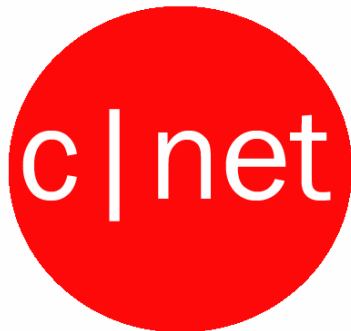
- Migrate mainframe application to the web
- Make accessible from 88 county courts

Solution

- SSL acceleration, compression, load balancing, and TCP/IP offload

Results:

- **Reduced bandwidth by 80%**
- **Eliminated need to build out additional T1s to each court house**
- **Cost savings >\$1M annually**



Challenge

- Scale to support an Alexa Top 100 web site
- Reduce maintenance costs

Solution

- Load balancing, TCP/IP offload

Results:

- Reduced power use by \$250k/yr
- Reduced server count
- Performance increase

Pacific Sunwear (Pacsun)



Challenge

- Improve shopping experience
- Reduce bandwidth requirements

Solution

- Load balancing, compression, TCP/IP offload, and SSL acceleration

Results:

- Bandwidth savings of 58%
- Cost savings of \$10k/month
- Marked revenue improvement



Challenge

- Stop server spiral under heavy load
- Consistent user experience

Solution

- SLB, compression, TCP/IP offload, SSL acceleration, and Surge Protection

Results:

- Ping pong through 4 hurricanes
- Eliminated need for massive rewrite
- Marked revenue improvement

Thank You!



Contact:

Steve Shah

Director of Product Management, Security Products
Citrix Systems, Inc.

Email: steve.shah@citrix.com