

Scaling Apache 2.x to > 20,000 users

colm.maccarthaigh@heanet.ie

colm@apache.org





Material

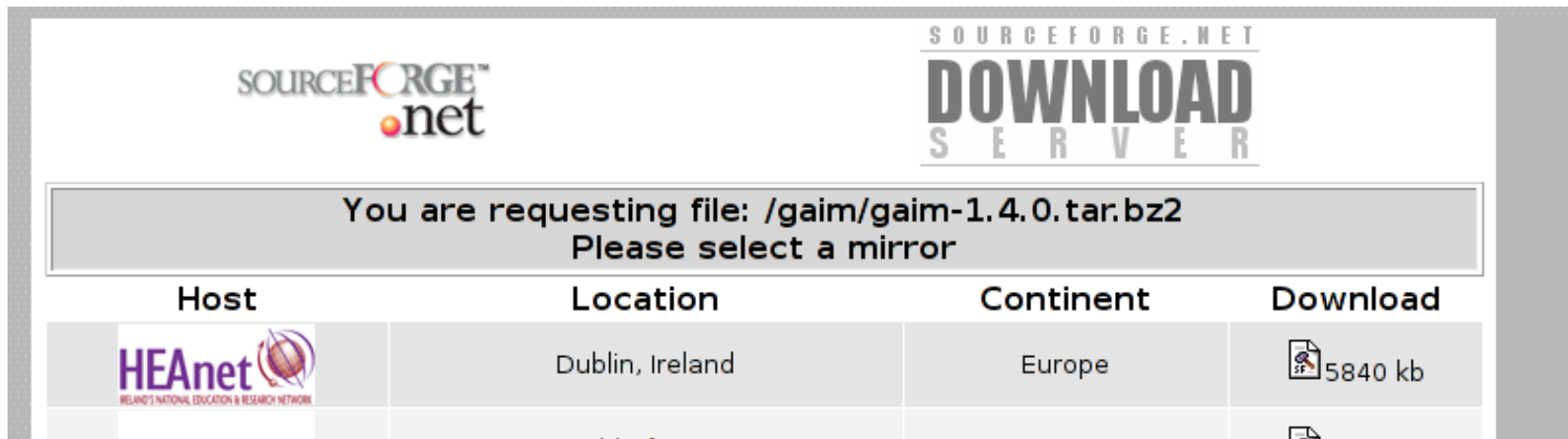
- Introduction
- Benchmarking
- Tuning Apache
- Tuning the Operating System
- Design of <ftp.heanet.ie>

ftp.heanet.ie



- National Mirror Server for Ireland
 - <http://ftp.heanet.ie/about/>
 - <http://ftp.heanet.ie/status/>
- Used for Network/Systems Development
 - IPv6, Jumboframes, Multicast, etc
 - Apache 2.0/2.1/2.2

Mirror for

- Apache, Sourceforge, Debian, FreeBSD, RedHat, Fedora, Slackware, Ubuntu, NASA Worldwinds, Mandrake, SuSe, Gentoo, Linux, OpenBSD, NetBSD ... and much much more.



The screenshot shows the SourceForge.net Download Server interface. At the top left is the SourceForge.net logo. At the top right is the text "SOURCEFORGE.NET" above "DOWNLOAD SERVER". A grey box in the center contains the text: "You are requesting file: /gaim/gaim-1.4.0.tar.bz2" and "Please select a mirror". Below this is a table with four columns: Host, Location, Continent, and Download. The first row shows the HEAnet mirror in Dublin, Ireland, Europe, with a file size of 5840 kb.

Host	Location	Continent	Download
	Dublin, Ireland	Europe	 5840 kb



The Numbers

- Roughly 27,000 concurrent downloads.
- 984 Mbit/sec in production.
- 4Gbit/sec in testing.
- Roughly 80% of all Sourceforge downloads from April 2004 to April 2004.
- Usually 4 times busier than ftp.kernel.org

The numbers: a day

- 11 million files stored
- 5 Terabytes of content available
- 3 million downloads
- 3.5 Terabytes of content shipped

Resources

- <http://www.kegel.com/c10k.html>
- <http://httpd.apache.org/>
- <http://www.csn.ul.ie/~mel/projects/vm/>
- Kernel sources
- Tuning/NFS/high-availability HOWTO's
- "Performance Tuning for Linux Servers"



Methodology

- Research the principles
- Configure, test, benchmark
- Configure, test, benchmark

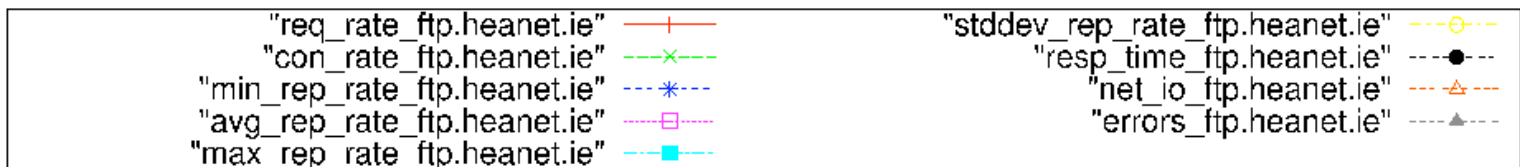
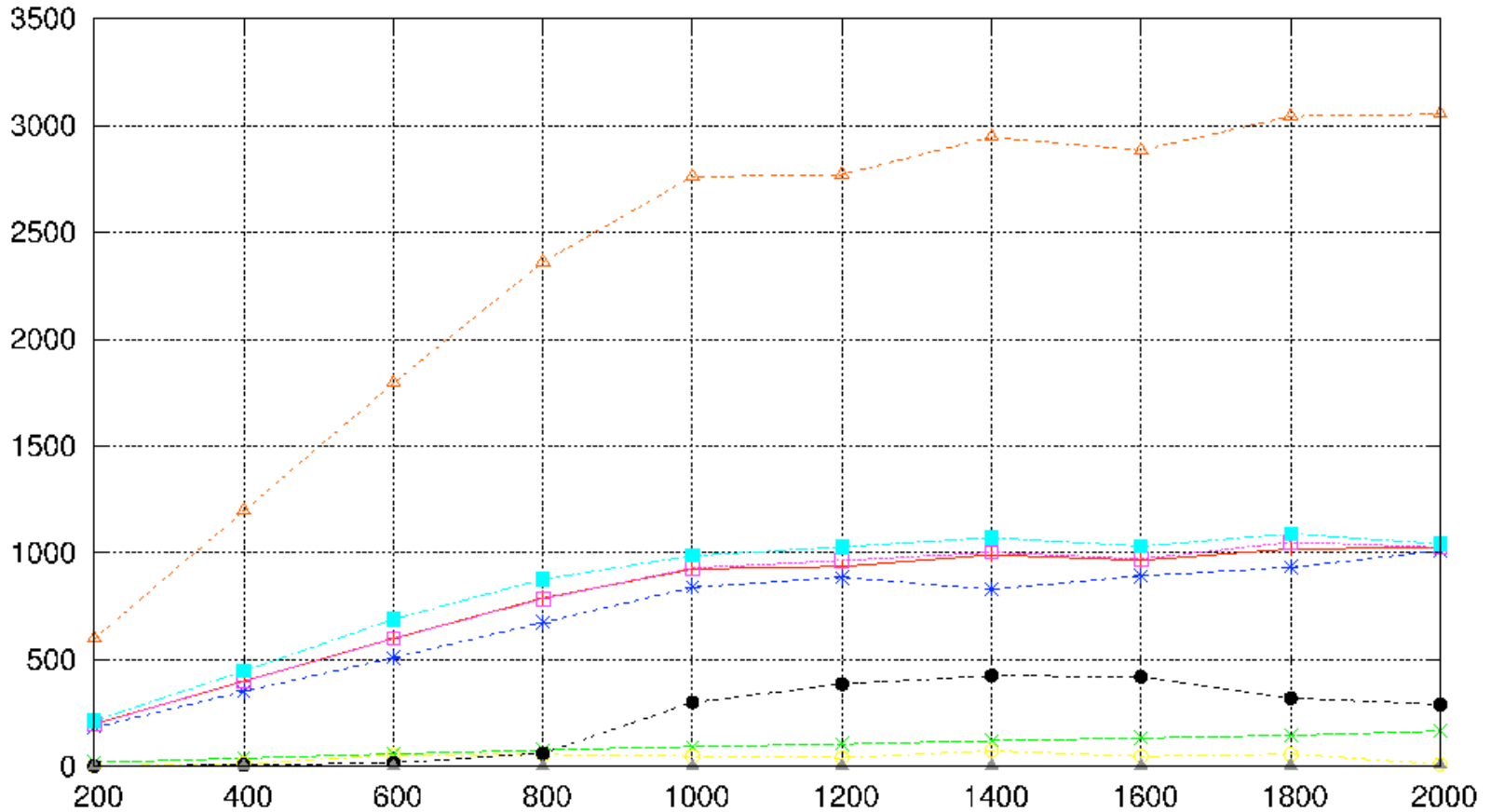
Benchmarking

- Webservers benchmarking:
 - apachebench, httpperf, autobench
- Always use the same files for benchmarking;
 - /ftp/pub/100.txt
 - /ftp/pub/1000.txt
 - /ftp/pub/10000.txt

Benchmarking

- Ab gives a good overview of webserver performance
- httpperf + autobench stress-tests and produces useful graphs, to visualise the maximum response rate, error rates, etc.

Without proxy



Benchmarking Filesystems

- IOZone, Postmark, Bonnie++
 - Postmark aimed at simulating mail-spools
 - IOZone is extensive and thorough
 - bonnie++ is simpler, and sufficient for most needs

Benchmarking the scheduler and VM

- No generic tools for for benchmarking schedulers or VMs
- Ad-hoc benchmarks usually consist of compiling a kernel
- We use `dder.sh`

```
#!/bin/sh
```

```
STARTNUM="1"
```

```
ENDNUM="102400"
```

```
# create a 100 MB file
```

```
dd bs=1024 count=102400 if=/dev/zero of=local.tmp
```

```
# Clear the record
```

```
rm -f record
```

```
# Find the most efficient size
```

```
for size in `seq $STARTNUM $ENDNUM`; do
```

```
    dd bs=$size if=local.tmp of=/dev/null 2>> record
```

```
done
```

```
# get rid of junk
```

```
grep "transf" record | awk '{ print $7 }' | cut -b 2- | cat -n | \
```

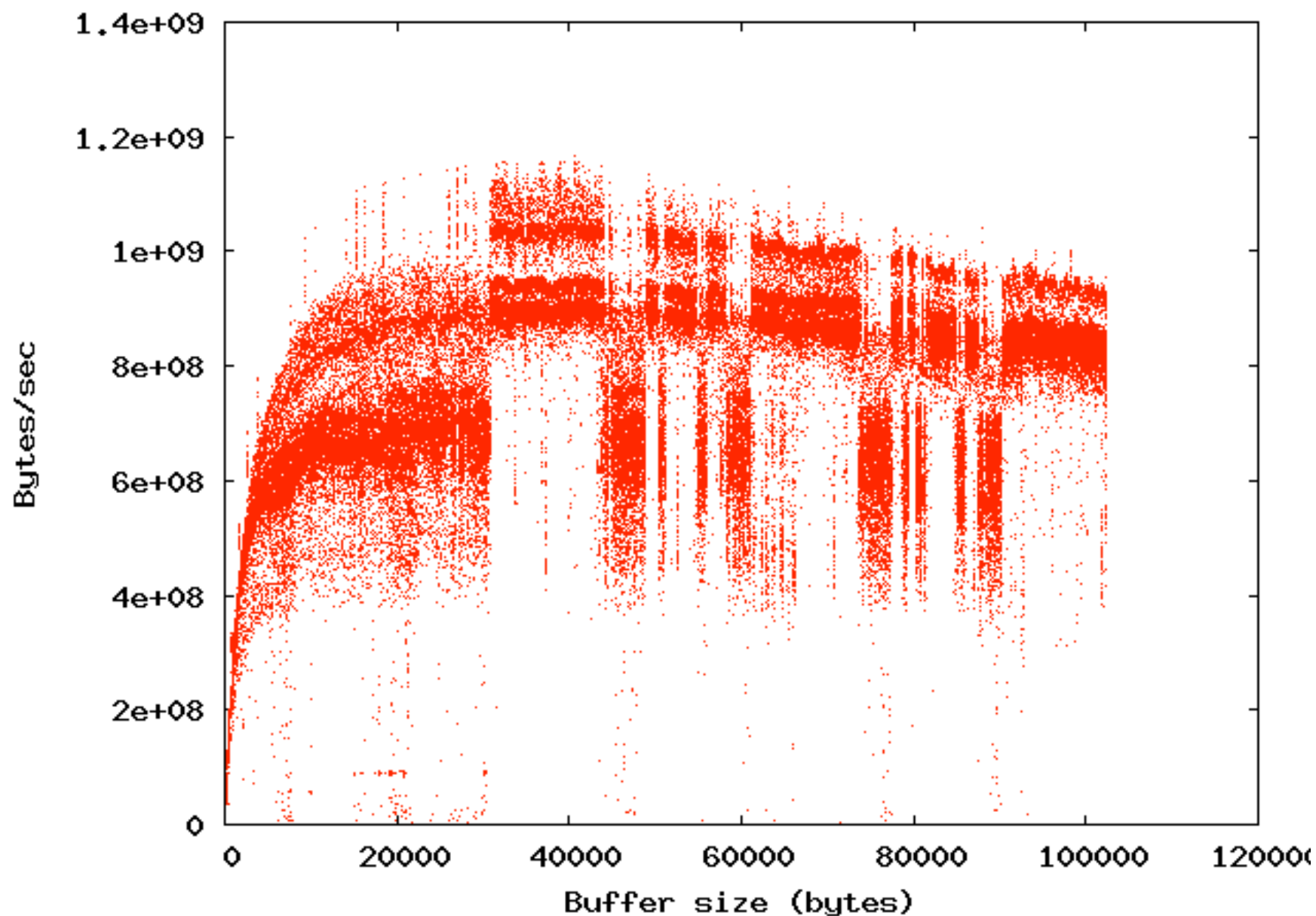
```
while read number result ; do
```

```
    echo -n $(( $number + $STARTNUM - 1 ))
```

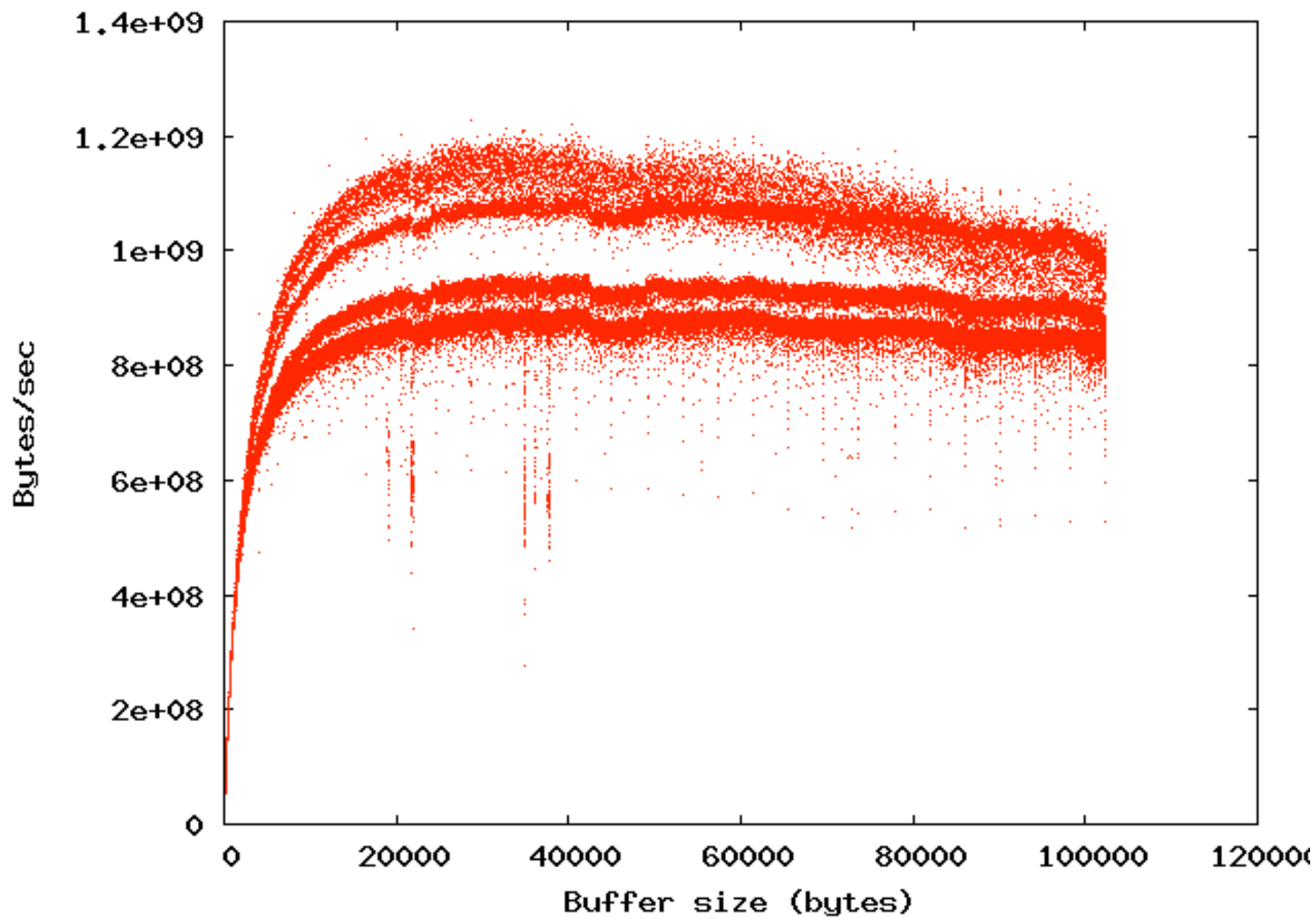
```
    echo " " $result
```

```
done > record.sane
```

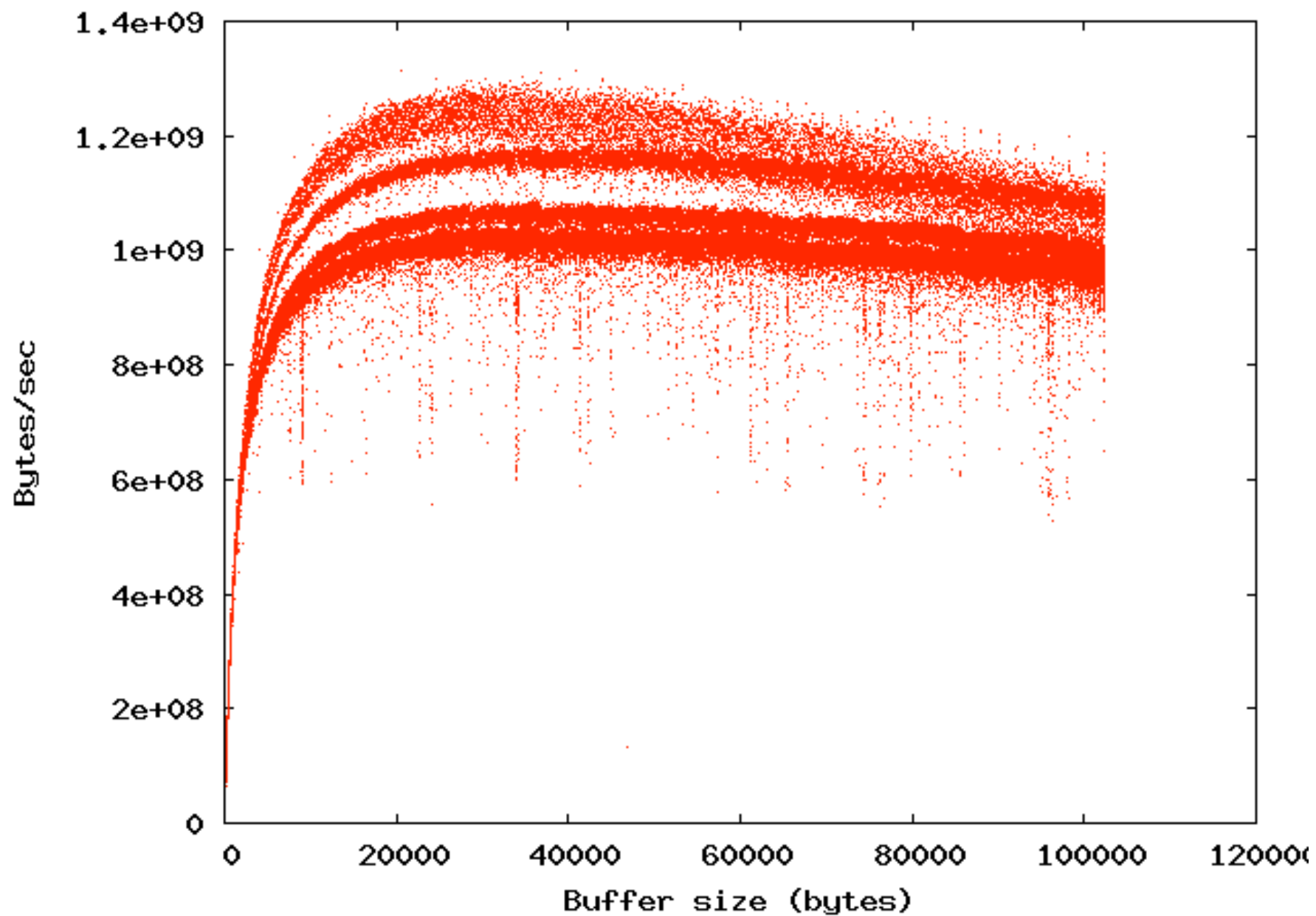
Buffer size efficiency (1Gb of RAM)



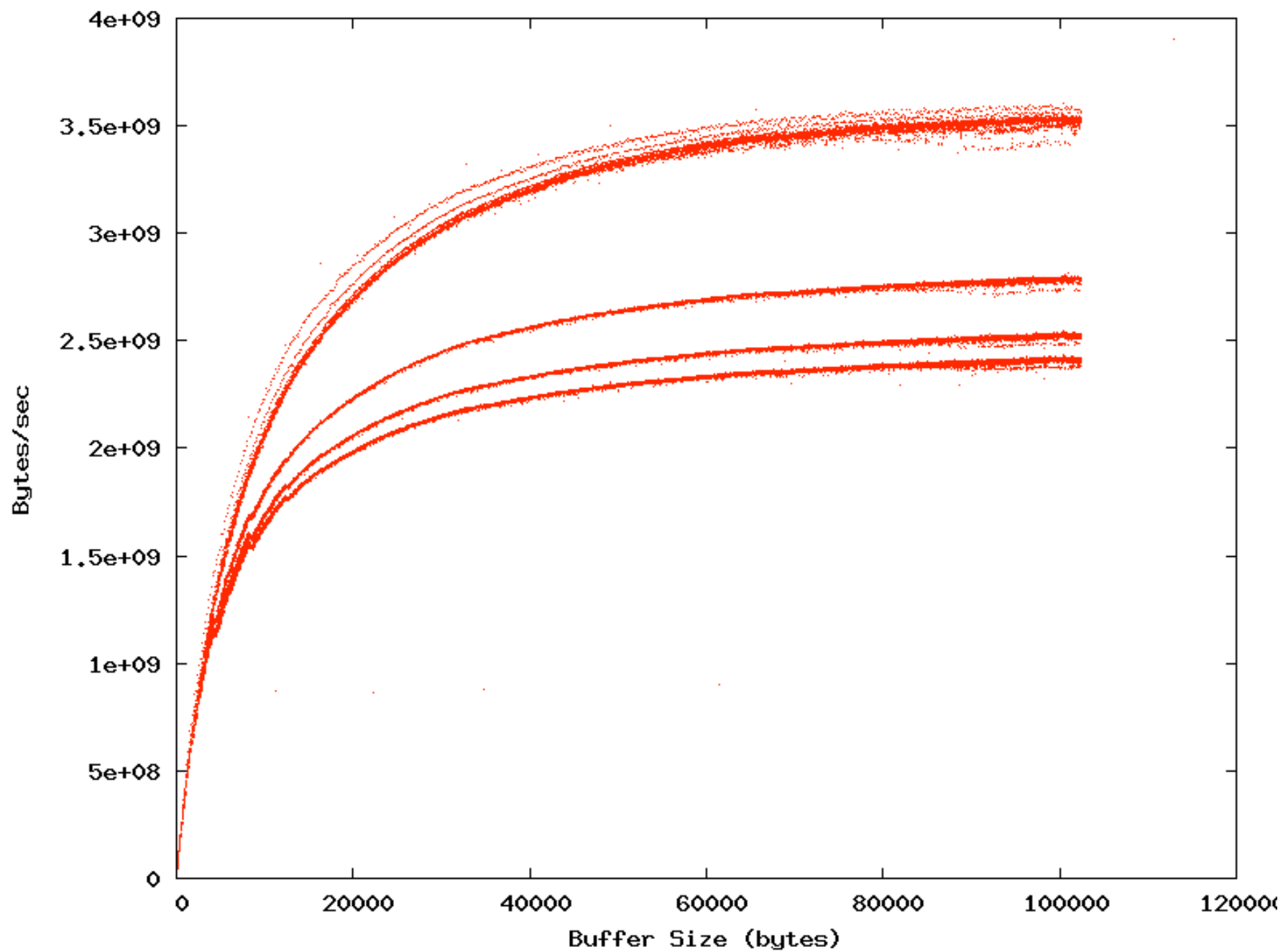
Buffer size efficiency (4Gb of RAM)



Buffer size efficiency (12Gb of RAM)



Buffer size efficiency (32Gb of RAM)



Tuning Apache

- Benchmark the MPMs
 - worker and event currently on top
- Static or DSO build for modules
 - miniscule difference
- AllowOverride none / EnableSendfile / EnableMMMap

mod_cache

- Experimental in 2.0, but usable in 2.1
- Not just for proxies, allows web servers to cache files as they are requested
- Many reads from a slow filesystem can be avoided

Tuning the Operating System

- Choose a kernel
 - 2.6 is much better than 2.4
- Tune the filesystem
 - always mount with noatime
 - XFS: use logbufs=8, ihashsize=65567
 - EXT3: set blocksize to 4096, use dir_index build option

Tuning NFS

- Use Jumboframes
 - increase wsize and rsize accordingly
- Increase the number of NFS threads on the server side
- Use nolock option if clients wont be writing



Sysctl

- vm/min_free_kbytes
- vm/lower_zone_protection
- vm/page-cluster
- vm/swappiness
- vm/vm_vfs_scan_ratio
- fs/file-max

Sysctl

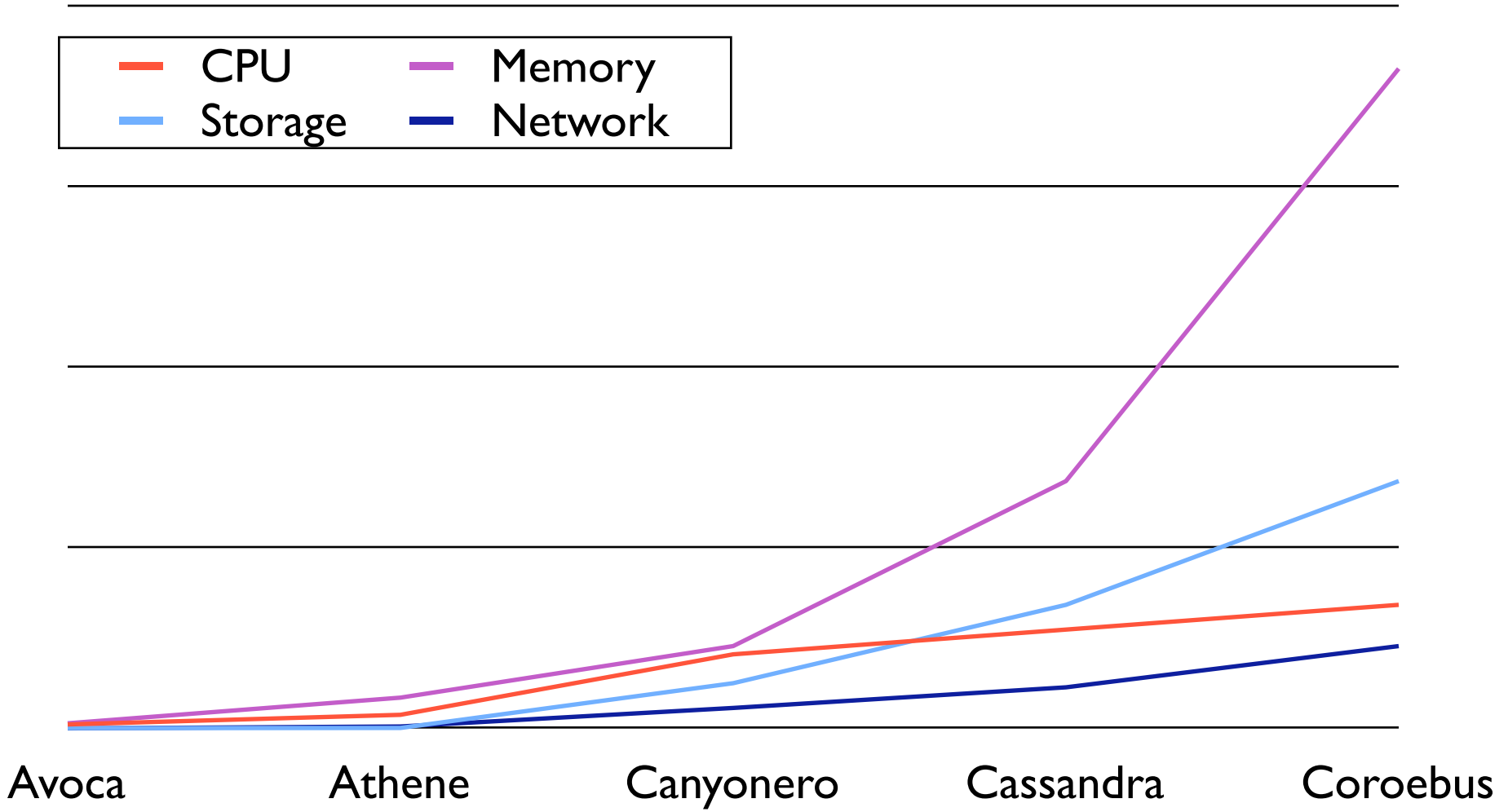
- net/ipv4/tcp_rfc1337
- net/ipv4/tcp_syncookies
- net/ipv4/tcp_keepalive_time
- net/ipv4/tcp_max_orphans
- sys/net/core/wmem_default
- sys/net/core/wmem_max

System Design

- Lots of Memory
- Bounce buffering and PAE to be avoided, otherwise lots of CPU
- Fast (15k RPM) SCSI disks for caching



Machine	Model	CPU	Memory	Storage	Network
Avoca	Alphaserver	200Mhz	256Mb	20Gb	10Mbit
Athene	Alphaserver DS20E	667Mhz	1.5Gb	30Gb	100Mbit
Canyonero	Dell 2650	2x1.8Ghz	4Gb	2.2Tb	1Gbit
Cassandra	Dell 2650	2x2.4Ghz	12Gb	5.6Tb	2Gbit
Coroebus	Dell 7250	2x1.5Ghz	32Gb	14.2Tb	4Gbit



Future services

- Multicast update notification
- mod_ftp for FTP content
- CVSup service



Questions

?