PLNOG

From the Earth to the Moon

From a Quagga-based Route Server to OpenBGPd

<elzbieta.jasinska@ams-ix.net>

PLNOG

Agenda

- Why Route Servers?
- What do Route Servers do?
- IX Requirements
- Current Implementations
- OpenBGPd Testing

amsix
amsterdam internet exchange

PLNOG

Agenda

- Why Route Servers?
- What do Route Servers do?
- IX Requirements
- Current Implementations
- OpenBGPd Testing

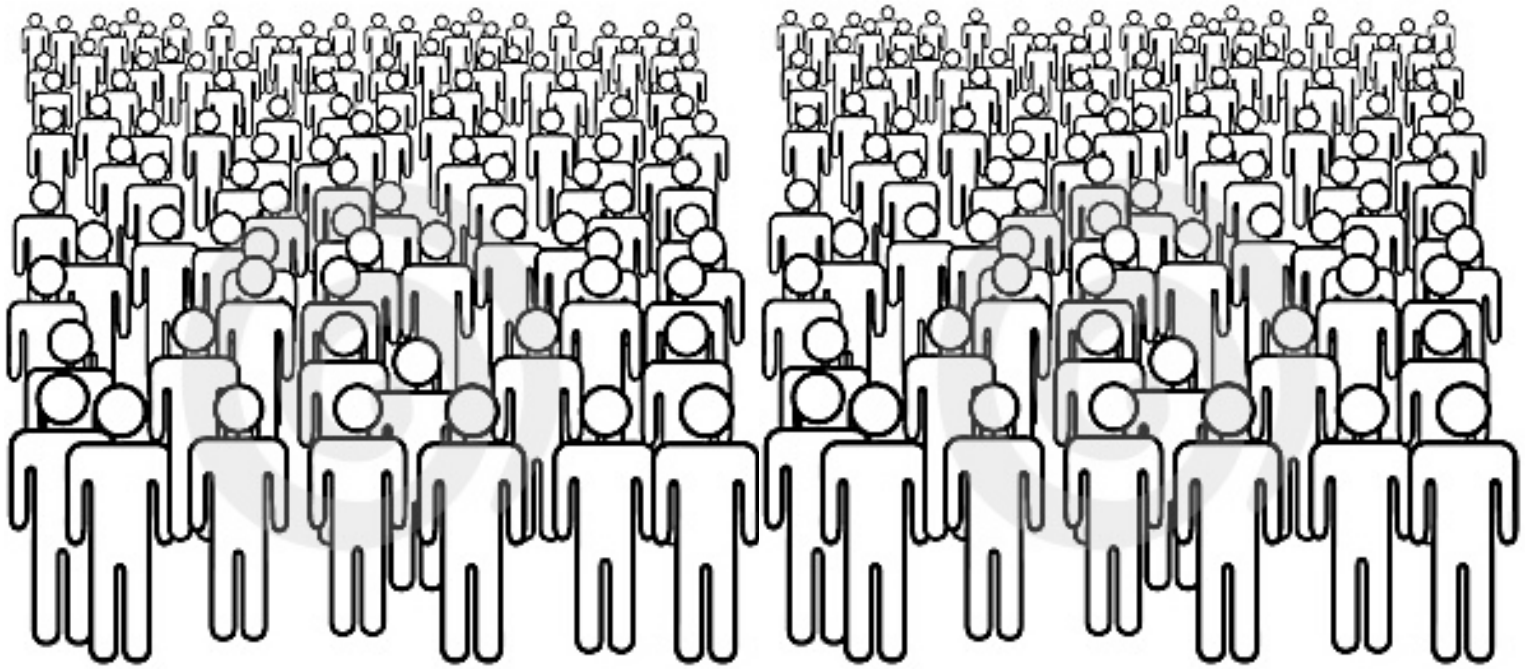
amsix
amsterdam internet exchange

Why Route Servers?

- Internet Exchange (e.g. AMS-IX)
- Peering platform for many parties
- Route Servers for the participants

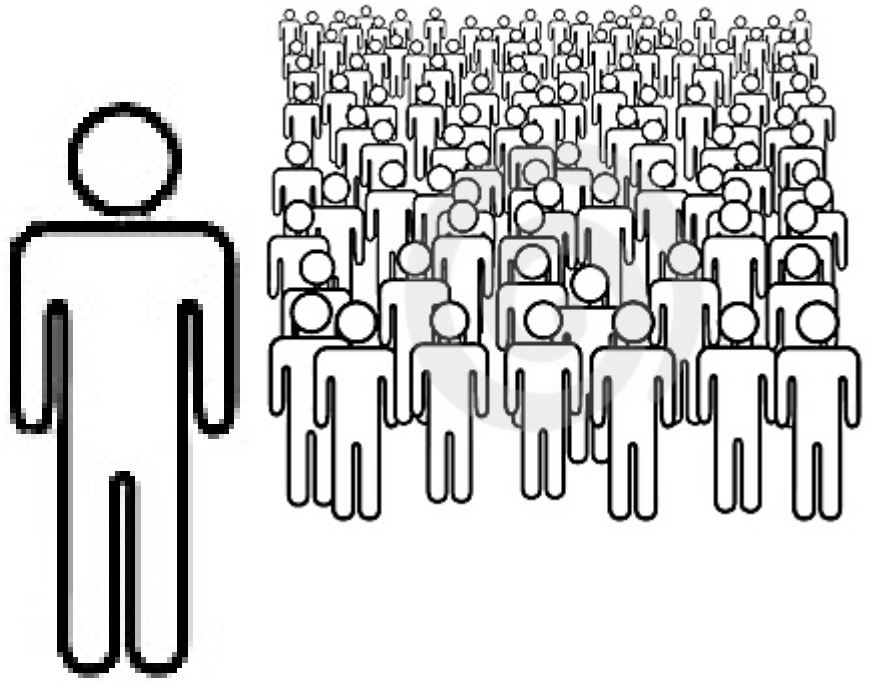
Why Route Servers?

- Peer with as many parties as possible
- ➡ Maintaining lots of BGP sessions



Why Route Servers?

- Reach a lot of parties with just one BGP session



PLNOG

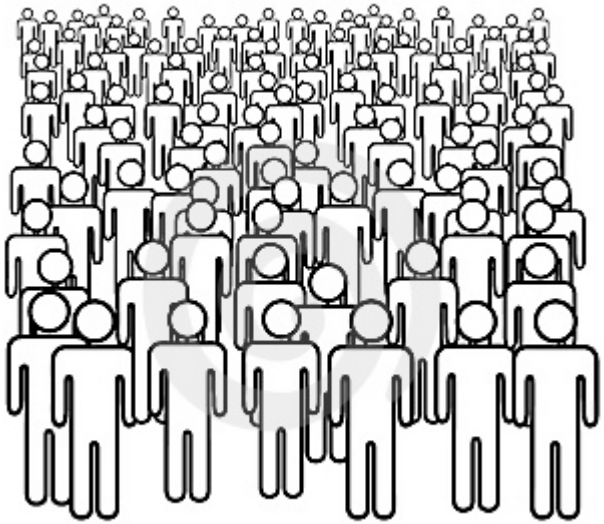
Why Route Servers?

- Redundancy ... in case your sessions die ...

amsix
amsterdam internet exchange

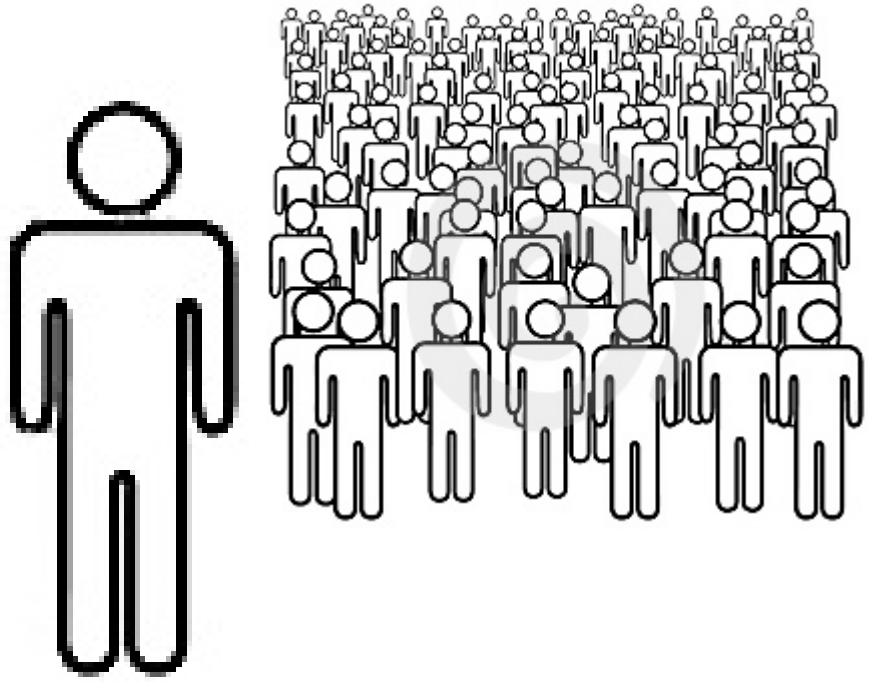
Why Route Servers?

- Redundancy ... in case the Route Server dies ...



Why Route Servers?

- Easy entry point for new members to the Exchange - immediate traffic



PLNOG

Agenda

- Why Route Servers?
- What do Route Servers do?
- IX Requirements
- Current Implementations
- OpenBGPd Testing

amsix
amsterdam internet exchange

What do route servers do?

- Receive UPDATEs from every participant

```
19:58:33.721679 IP (tos 0x0, ttl 64, id 44576, offset 0, flags [DF], proto TCP (6),
length 117) 10.23.0.5.58880 > 10.23.0.1.179: P, cksum 0xb892 (correct), 48:113(65) ack
61 win 1460 <nop,nop,timestamp 1976783474 3177206804>: BGP, length: 65
        Update Message (2), length: 65
...
          AS Path (2), length: 10, Flags [T]: 65499 11 12 13
...
          Next Hop (3), length: 4, Flags [T]: 10.23.0.5
...
          Updated routes:
            2.0.5.0/24
19:58:33.723897 IP (tos 0x0, ttl 64, id 42762, offset 0, flags [DF], proto TCP (6),
length 117) 10.23.0.4.33349 > 10.23.0.1.179: P, cksum 0xb033 (correct), 48:113(65) ack
61 win 1460 <nop,nop,timestamp 1976783474 1916183085>: BGP, length: 65
        Update Message (2), length: 65
...
          AS Path (2), length: 10, Flags [T]: 65500 11 12 13
...
          Next Hop (3), length: 4, Flags [T]: 10.23.0.4
...
          Updated routes:
            2.0.4.0/24
```

What do route servers do?

- Apply filters for the receiving peers

```
from AS65500 accept ANY
to AS65500 announce AS65499

from AS65499 accept ANY
to AS65499 announce AS65500
```

What do route servers do?

- Perform "best path" selection for every peer
- Store Routing Information Base (RIB) for every peer

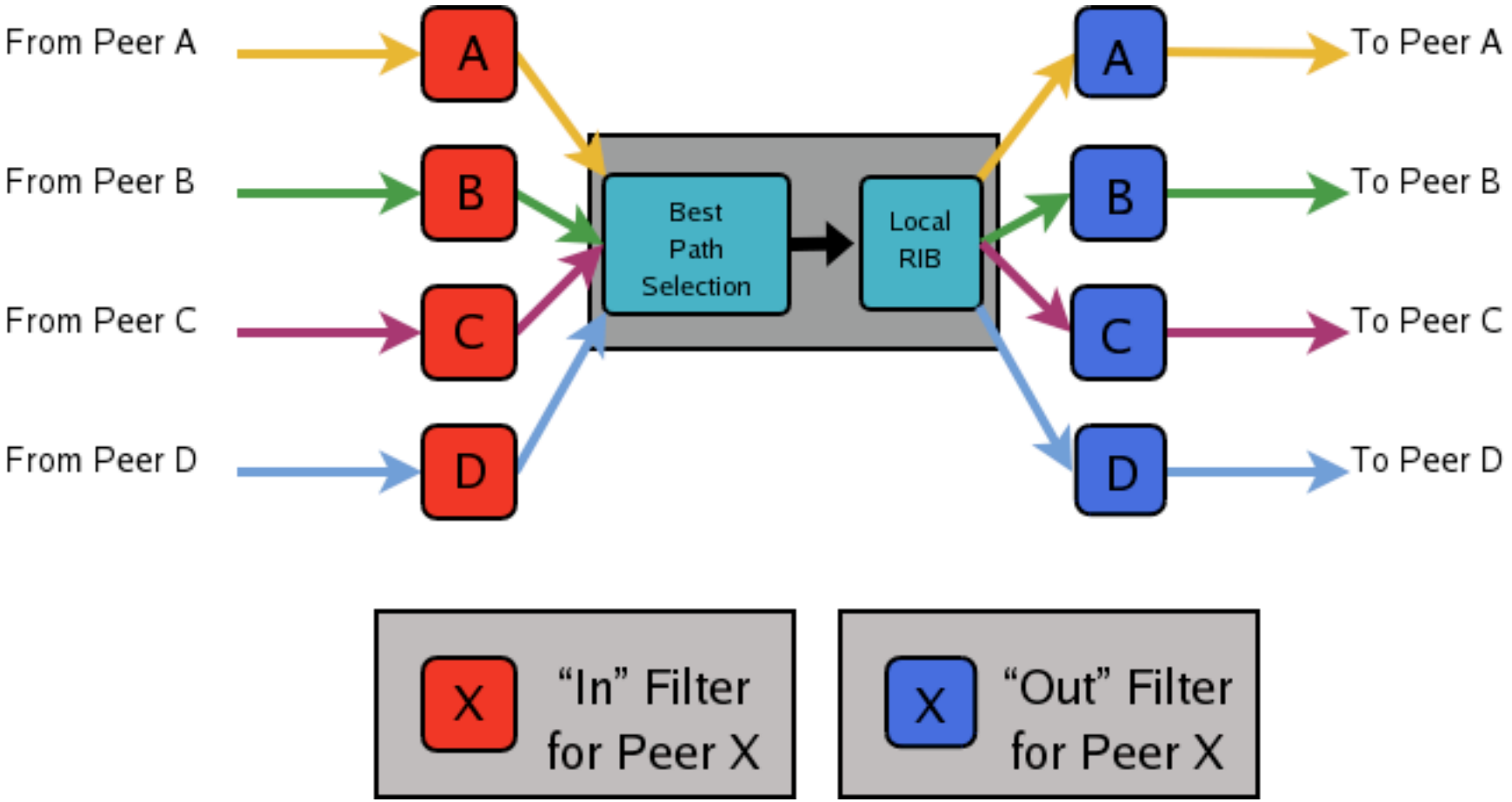
```
flags destination          gateway          lpref   med aspath origin
*>    2.0.4.0/24           10.23.0.4          100   200 65500 11 12 13 i
*>    2.0.5.0/24           10.23.0.5          100   200 65499 11 12 13 i
```

What do route servers do?

- Forward the RIB contents to the desired peer

```
19:58:33.901718 IP (tos 0xc0, ttl 1, id 15745, offset 0, flags [DF], proto TCP (6),
length 103) 10.23.0.1.179 > 10.23.0.4.33349: P, cksum 0x2b21 (correct), 61:112(51) ack
114 win 17376 <nop,nop,timestamp 1916183105 1976783474>: BGP, length: 51
        Update Message (2), length: 51
...
          AS Path (2), length: 10, Flags [T]: 65499 11 12 13
...
          Next Hop (3), length: 4, Flags [T]: 10.23.0.5
...
          Updated routes:
            2.0.5.0/24
19:58:33.903463 IP (tos 0xc0, ttl 1, id 12268, offset 0, flags [DF], proto TCP (6),
length 103) 10.23.0.1.179 > 10.23.0.5.58880: P, cksum 0x377e (correct), 61:112(51) ack
114 win 17376 <nop,nop,timestamp 3177206824 1976783474>: BGP, length: 51
        Update Message (2), length: 51
...
          AS Path (2), length: 10, Flags [T]: 65500 11 12 13
...
          Next Hop (3), length: 4, Flags [T]: 10.23.0.4
...
          Updated routes:
            2.0.4.0/24
```

What do route servers do?



Source: Quagga Route Server model description

PLNOG

Agenda

- Why Route Servers?
- What do Route Servers do?
- IX Requirements
- Current Implementations
- OpenBGPd Testing

amsix
amsterdam internet exchange

PLNOG

Requirements

- Stability
- Performance
- Stability
- Performance
- Stability
- Performance

PLNOG

Requirements

- Allow for peering policies to be maintained
- Stability
- Performance
- Stability
- Performance
- ...

PLNOG

Agenda

- Why Route Servers?
- What do Route Servers do?
- IX Requirements
- Current Implementations
- OpenBGPd Testing

amsix
amsterdam internet exchange

Current Implementations

- Quagga
- OpenBGPd
- BIRD

Quagga

- Two Quagga instances currently in use at AMS-IX
  - Each ~ 200 BGP sessions, announcing ~ 25.000 prefixes
- Single threaded implementation
  - Issues with performing its tasks on time

PLNOG

OpenBGPd

- Multi threaded implementation
  - ... more details in a moment

BIRD

- Mentioned for the sake of completeness
- ... research on the TODO list ...

PLNOG

Agenda

- Why Route Servers?
- What do Route Servers do?
- IX Requirements
- Current Implementations
- OpenBGPd Testing

amsix
amsterdam internet exchange

PLNOG

OpenBGPd Testing

"I don't know... A proof is a proof. What kind of a proof? It's a proof. A proof is a proof, and when you have a good proof, it's because it's proven."

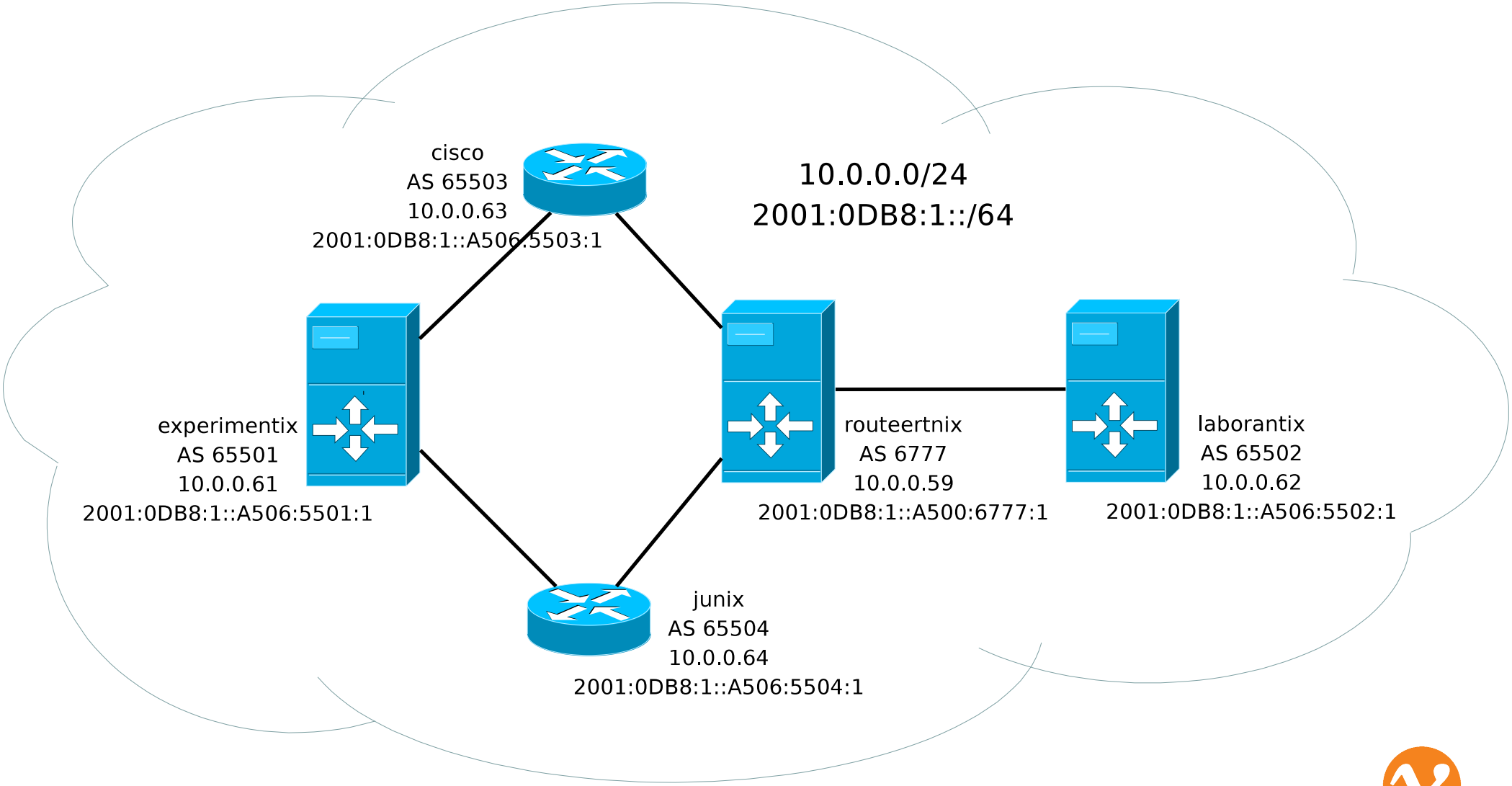
-- Jean Cheretien

Thanks to
Arnoud Vermeer and Leon Weber

OpenBGPd Testing

- Basic functionality
  - BGP protocol
  - Route selection
- Stress testing / performance
  - Amount of peering sessions
  - Amount of prefixes
  - Amount of UPDATEs

Test Setup #1



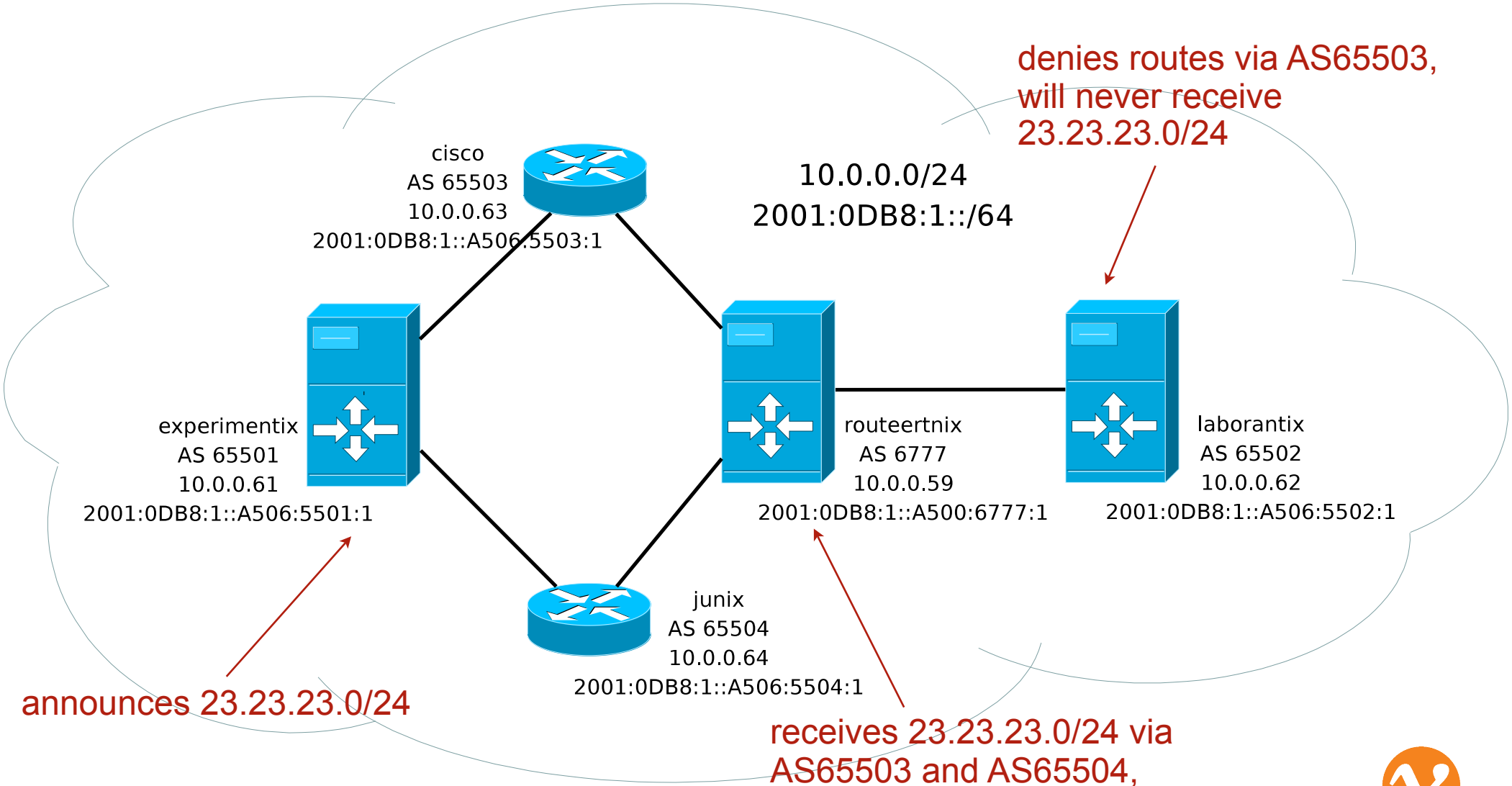
IPv6 Bug

- Upon "clearing" one peers session, other IPv6 sessions died
- It took some time to realize it had to do with withdraw messages

```
Mar 10 04:41:45 routeertnix bgpd[25100]: neighbor 2001:db8:1::a506:5502:1
(laborantix v6) AS65502: withdraw 2001:db8:97::/64
Mar 10 04:41:45 routeertnix bgpd[12120]: neighbor 2001:db8:1::a506:5504:1
(junix v6): received notification: error in UPDATE message, attribute list error
Mar 10 04:41:45 routeertnix bgpd[12120]: neighbor 2001:db8:1::a506:5504:1
(junix v6): state change Established -> Idle, reason: NOTIFICATION received
```

- IPv6 withdraws in OpenBGPd were erroneous
- Fixed since March 11th 2009

Per-Peer RIBs



Per-Peer RIBs

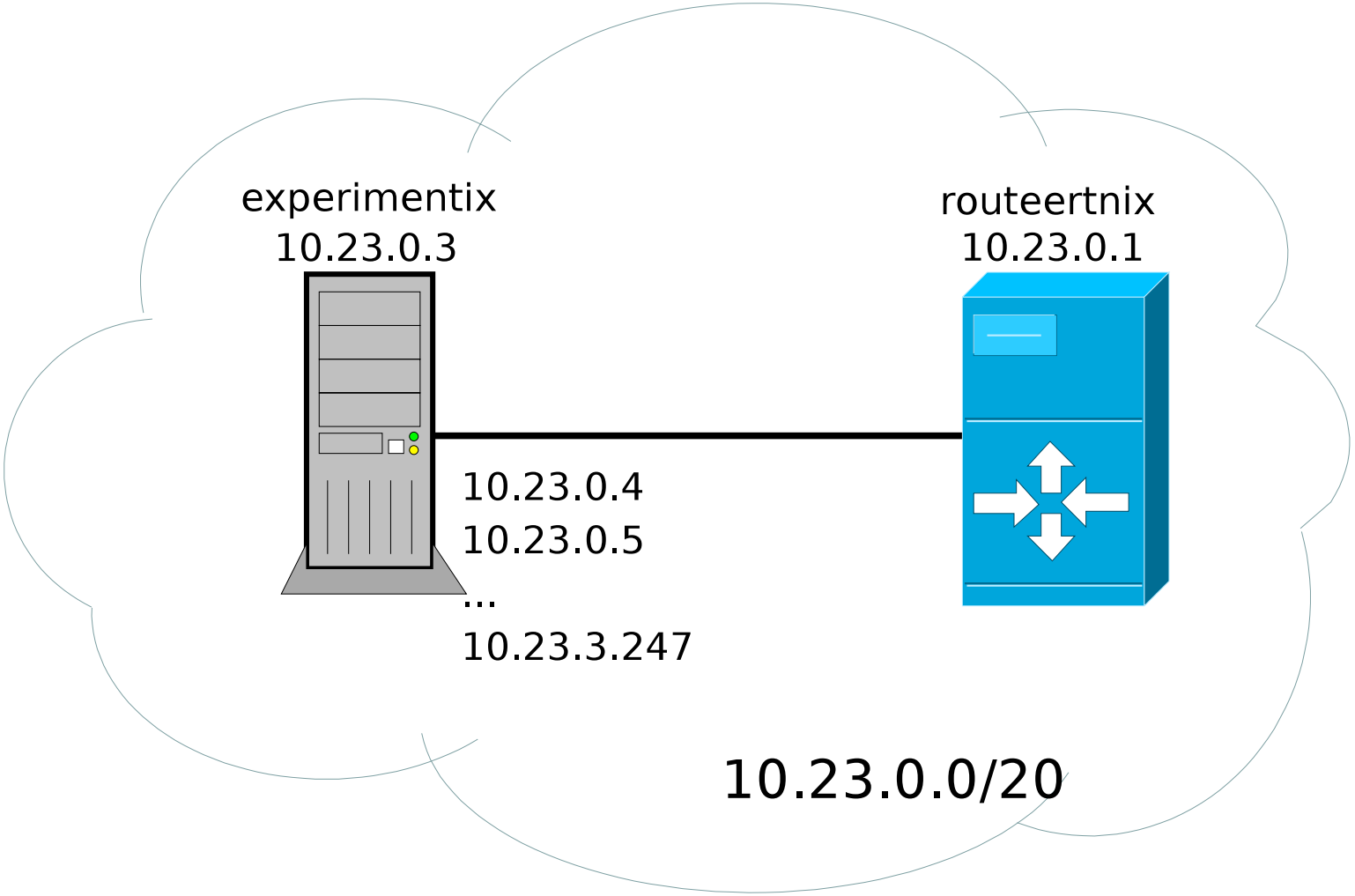
- OpenBGPd so far only provided a global "best path" selection and RIB
- Client side filtering would not be applicable

- Per-Peer configurable RIBs have been implemented on request as of June 8th 2009

Test Setup #2

- Test Server - Experimentix
  - Directly connected to the Route Server
  - Lots of IP aliases on its interface
  - Runs a Perl script based on Net::BGP to simulate BGP speakers (http://www.ams-ix.net/downloads/)

Test Setup #2



Open File Limit

- Establishing over 1024 BGP sessions to the Route Server caused:

```
Sep  8 22:04:01 routeertnix bgpd[19382]: connection from non-peer (unknown) refused
Sep  8 22:04:01 routeertnix bgpd[19382]: accept: Too many open files
```

- Default open file limit on OpenBSD: 1024

```
# ulimit -a
...
open files                      (-n) 1024
...
```

- Starting the daemon with an increased "ulimit -n" helps

```
(ulimit -n 128000; bgpd -vv)
```

First UPDATEs

- We were ready to go...
  - Setup was prepared
  - Sessions were coming UP
  - Different bgpd.conf files were prepared
  - Graphing scripts were written
    to plot CPU/Mem usage with different configs

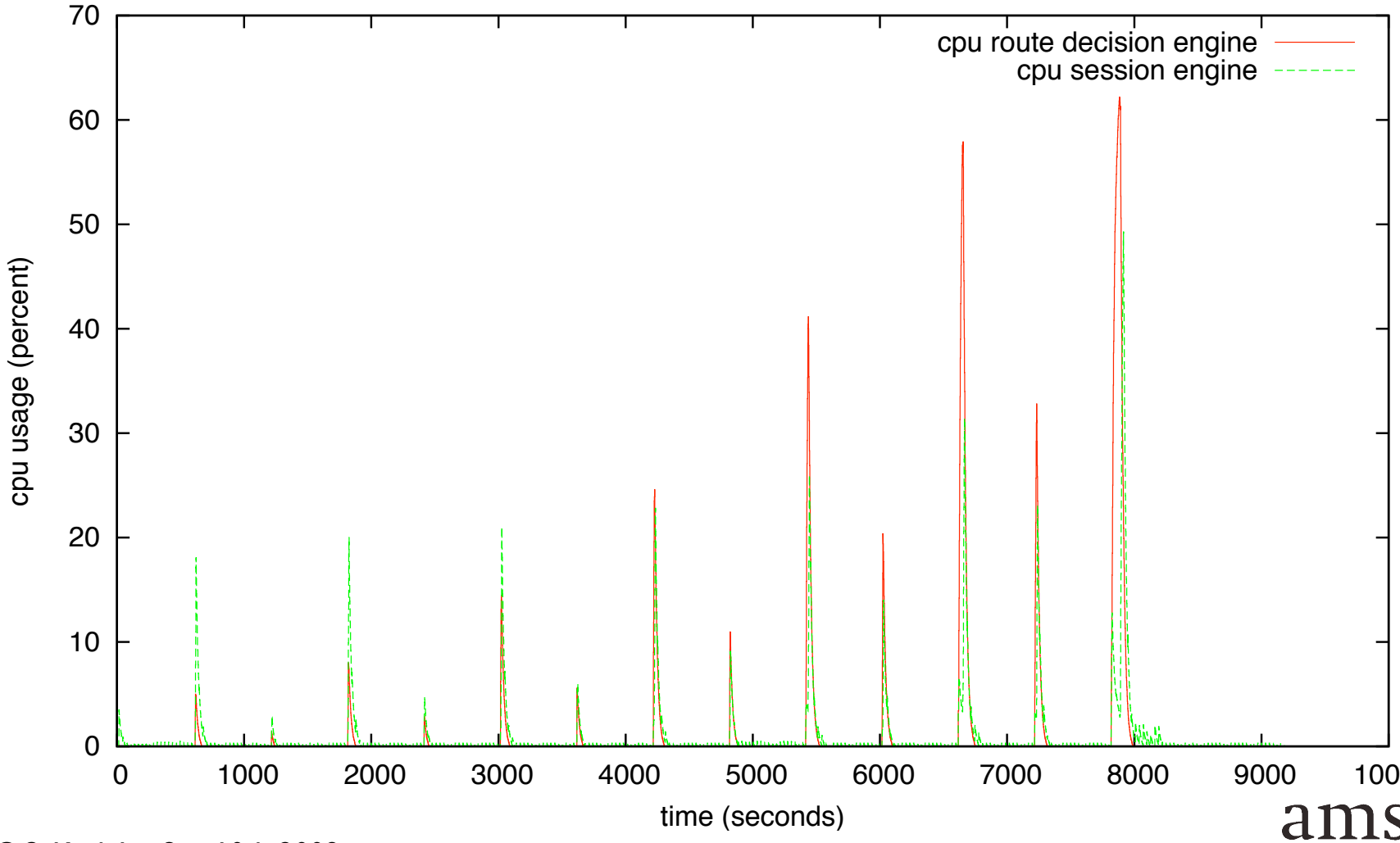
PLNOG

Test Sequence

- Increase # of sessions: 253, 506, 1012 per test
- Test single and multiple RIBs
- Increase # of unique announcements/ withdraws: 1,2,4,8,16,32,64 within one test, separated by a defined interval

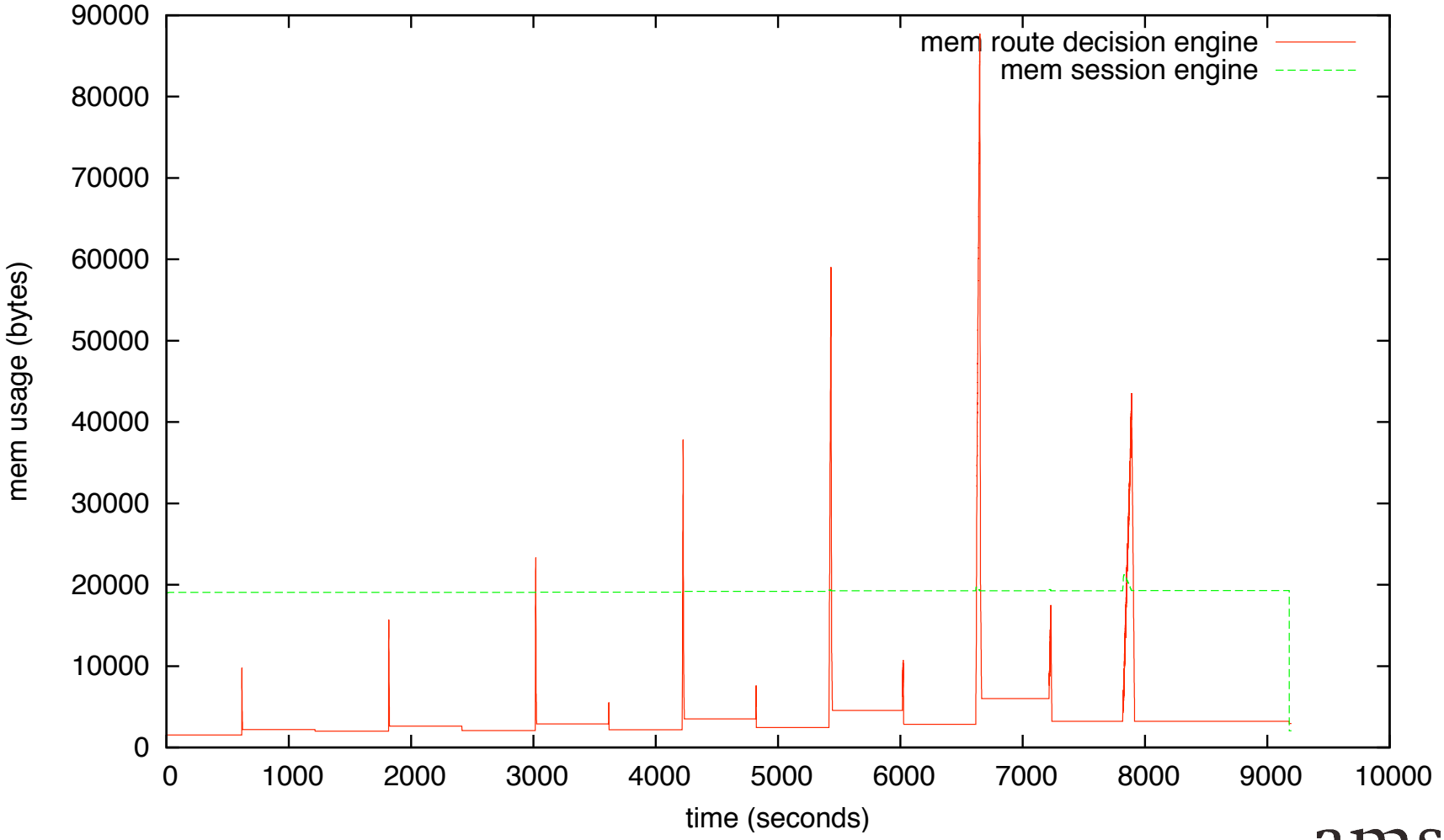
Test #1 - CPU

OpenBGPd cpu usage
single-rib; 253 sessions established at t=0
1/2/4/8/16/32/64 updates per sess every 10min
announce/withdraw 1 unique prefix each
Intel(R) Pentium(R) III CPU family 1133MHz (GenuineIntel 686-class)



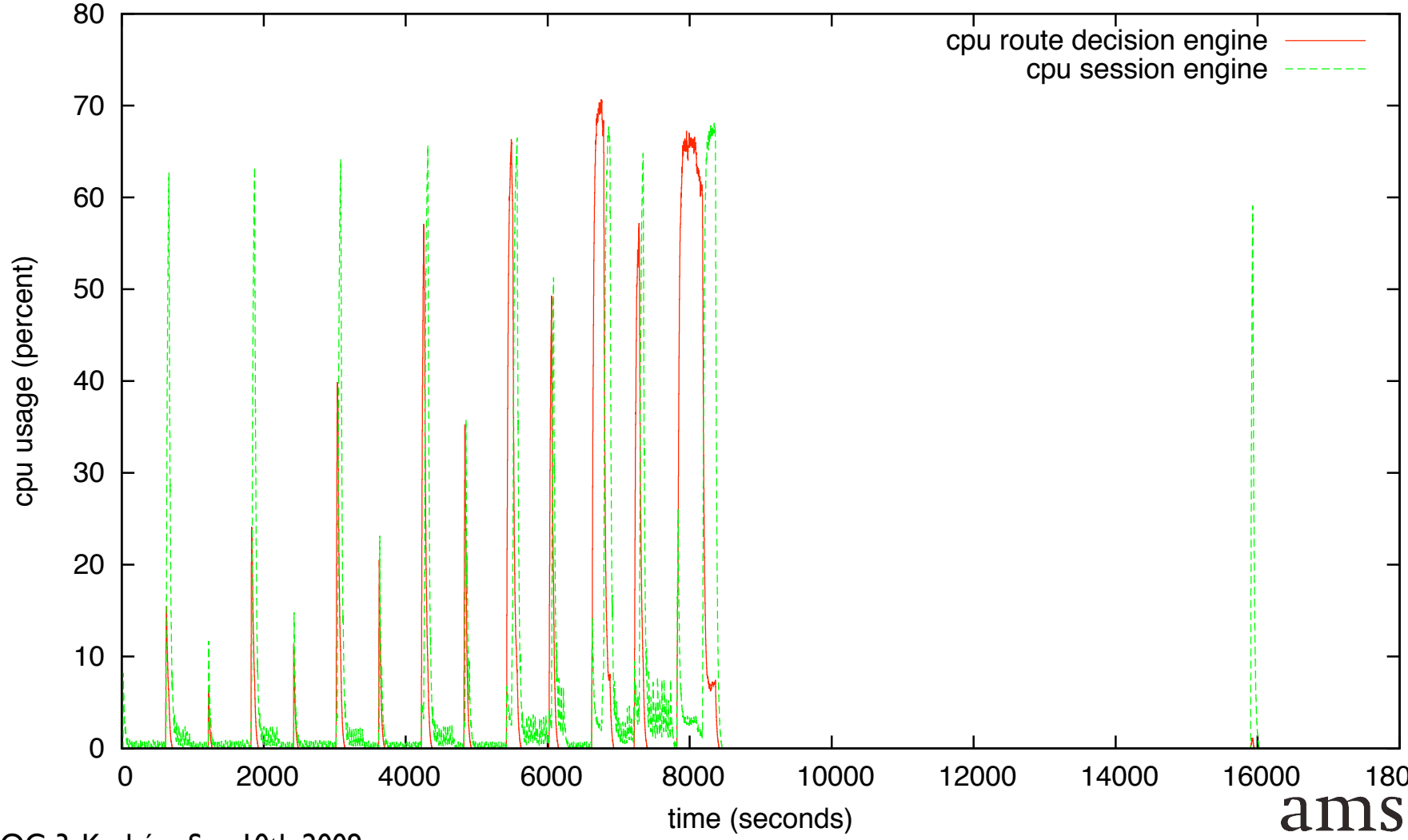
Test #1 - Mem

OpenBGPd mem usage
single-rib; 253 sessions established at t=0
1/2/4/8/16/32/64 updates per sess every 10min
announce/withdraw 1 unique prefix each
Intel(R) Pentium(R) III CPU family 1133MHz (GenuineIntel 686-class)



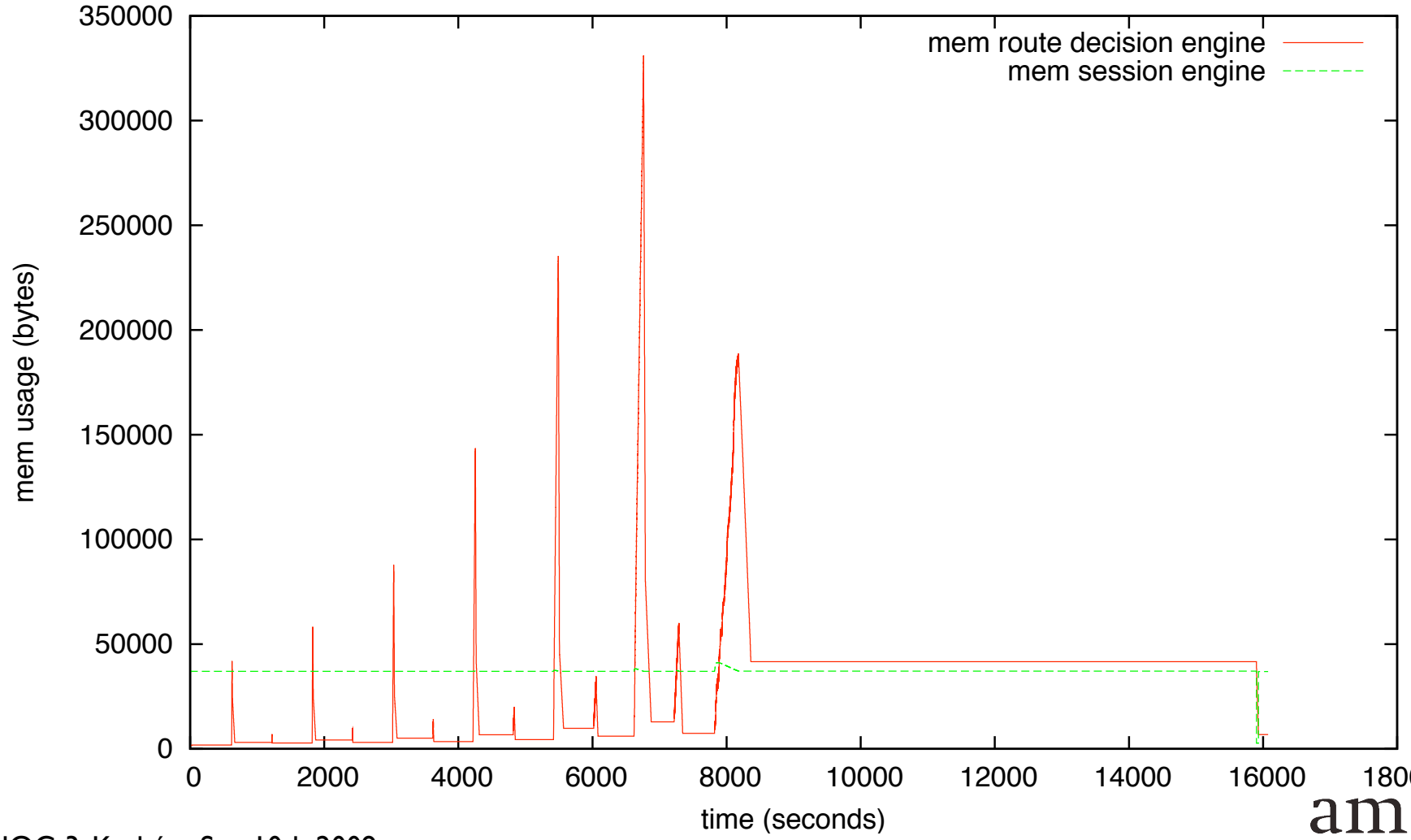
Test #2 - CPU

OpenBGPd cpu usage
single-rib; 506 sessions established at t=0
1/2/4/8/16/32/64 updates/withdraws per sess every 10min
announce 1 unique prefix each
Intel(R) Pentium(R) III CPU family 1133MHz (GenuineIntel 686-class)



Test #2 - Mem

OpenBGPd mem usage
single-rib; 506 sessions established at t=0
1/2/4/8/16/32/64 updates/withdraws per sess every 10min
announce 1 unique prefix each
Intel(R) Pentium(R) III CPU family 1133MHz (GenuineIntel 686-class)



Mbufs

- Suddenly the Route Server became unreachable...

```
WARNING: mclpool limit reached; increase kern.maxclusters
```

- Defaults

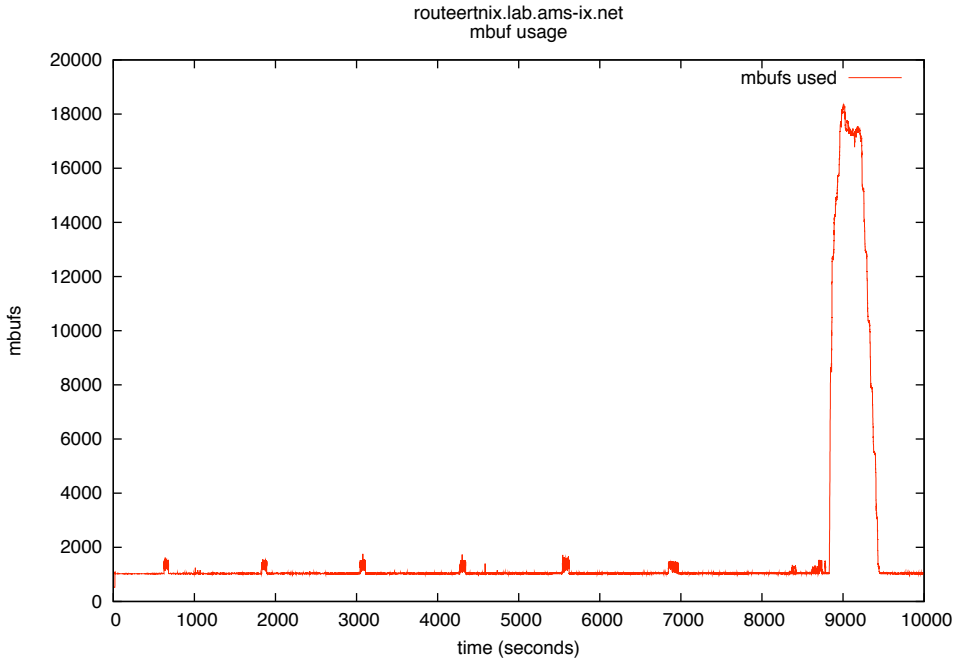
```
# sysctl kern.maxclusters
kern.maxclusters=6144
```

- Change them

```
# sysctl kern.maxclusters=240000
kern.maxclusters: 6144 -> 240000
```

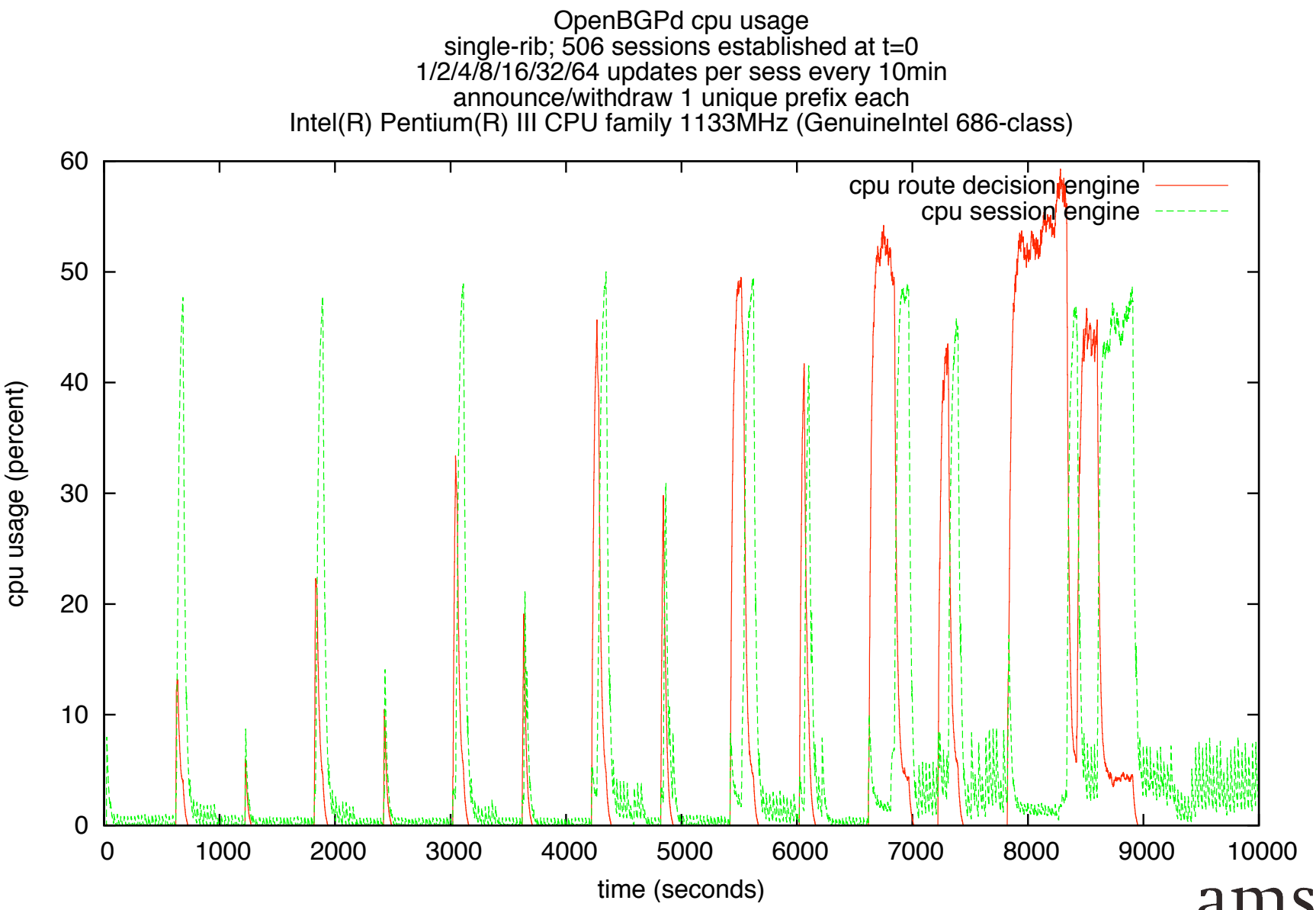
Mbufs

- We are now also plotting mbuf usage

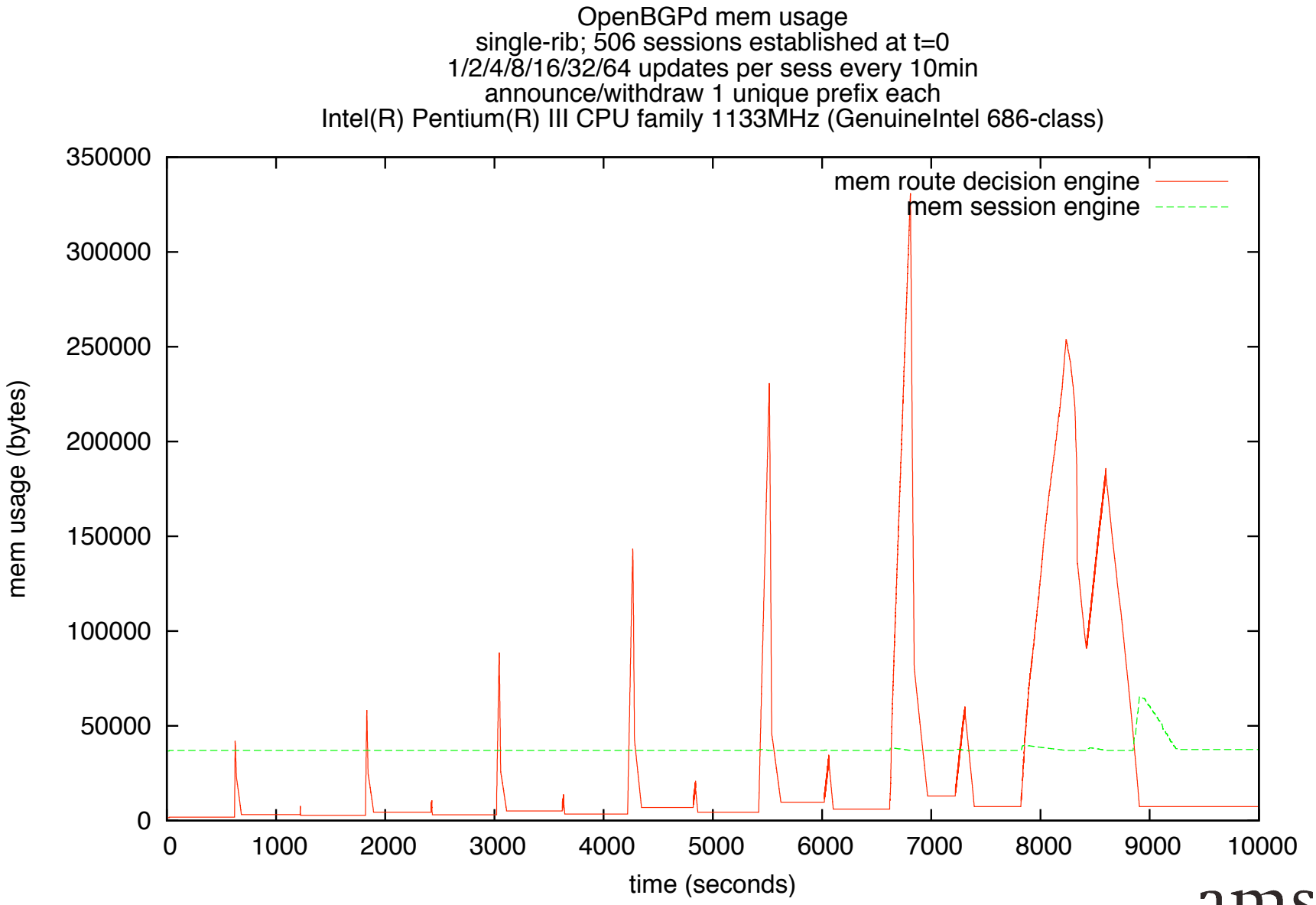


- Buffer usage peaks are still being looked into by the developers

Test #2 again - CPU

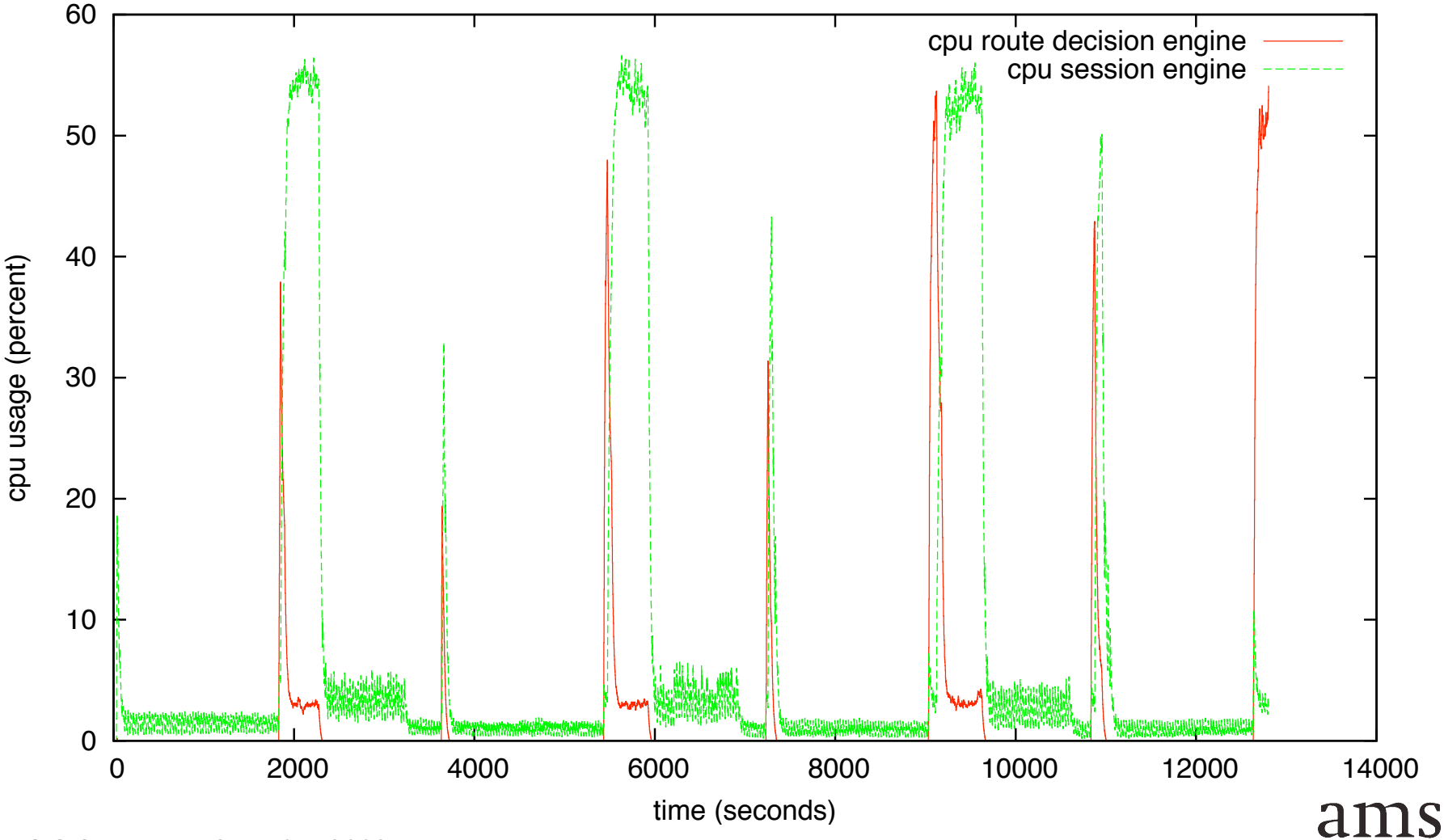


Test #2 again - Mem



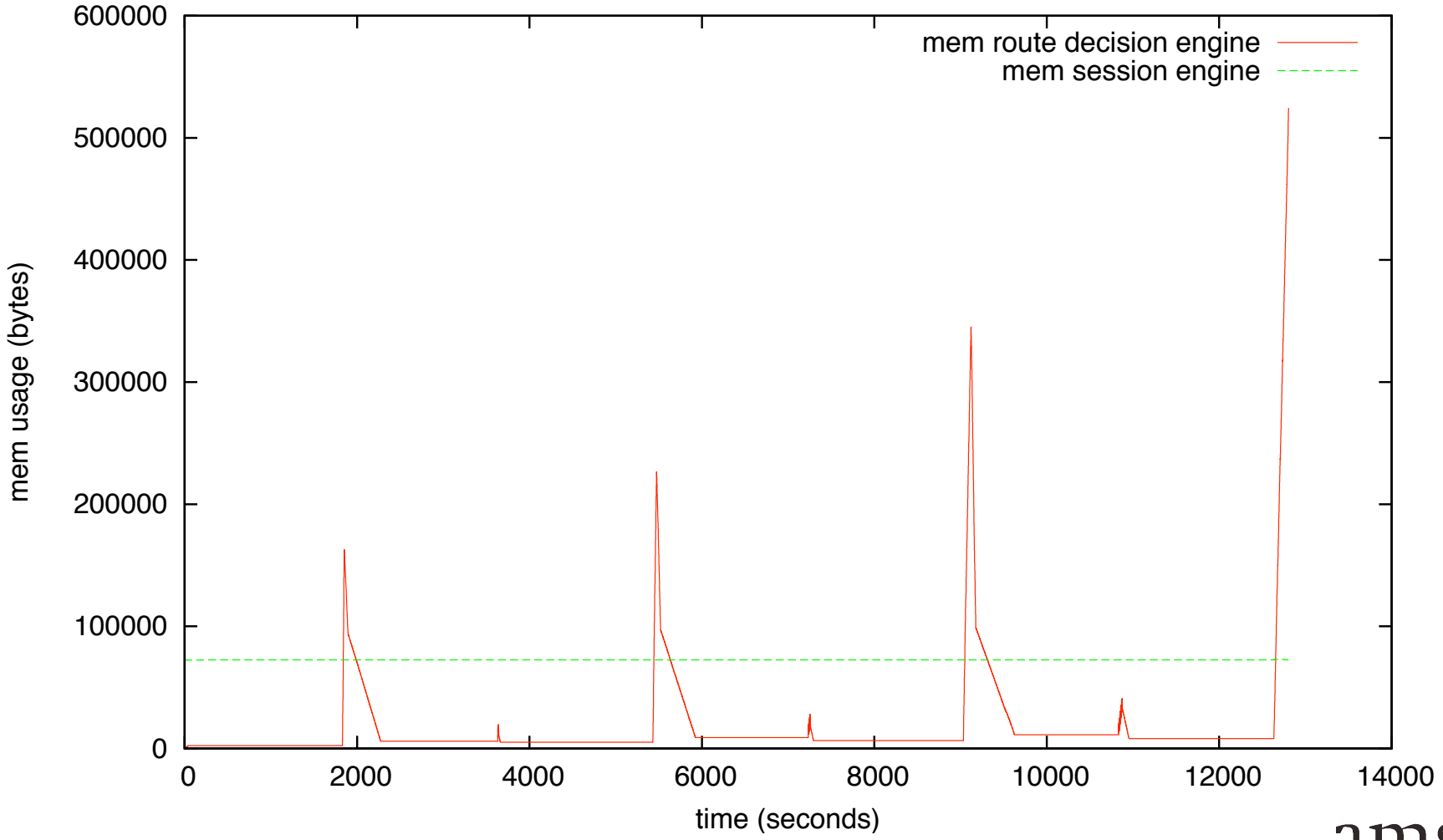
Test #3 - CPU

OpenBGPd cpu usage
single-rib; 1012 sessions established at t=0
1/2/4/8/16/32/64 updates per sess every 30min
announce/withdraw 1 unique prefix each
Intel(R) Pentium(R) III CPU family 1133MHz (GenuineIntel 686-class)



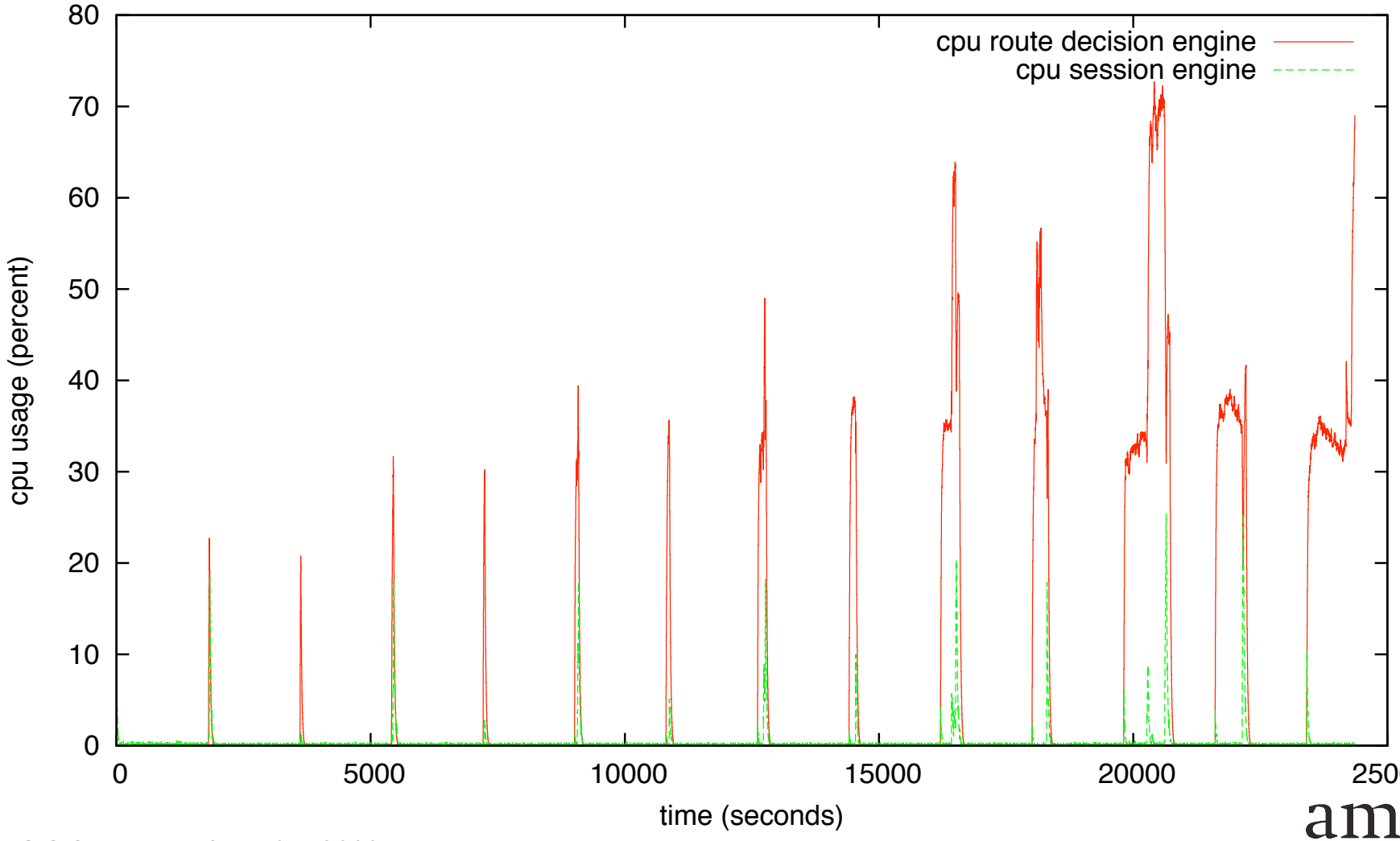
Test #3 - Mem

OpenBGPd mem usage
single-rib; 1012 sessions established at t=0
1/2/4/8/16/32/64 updates per sess every 30min
announce/withdraw 1 unique prefix each
Intel(R) Pentium(R) III CPU family 1133MHz (GenuineIntel 686-class)



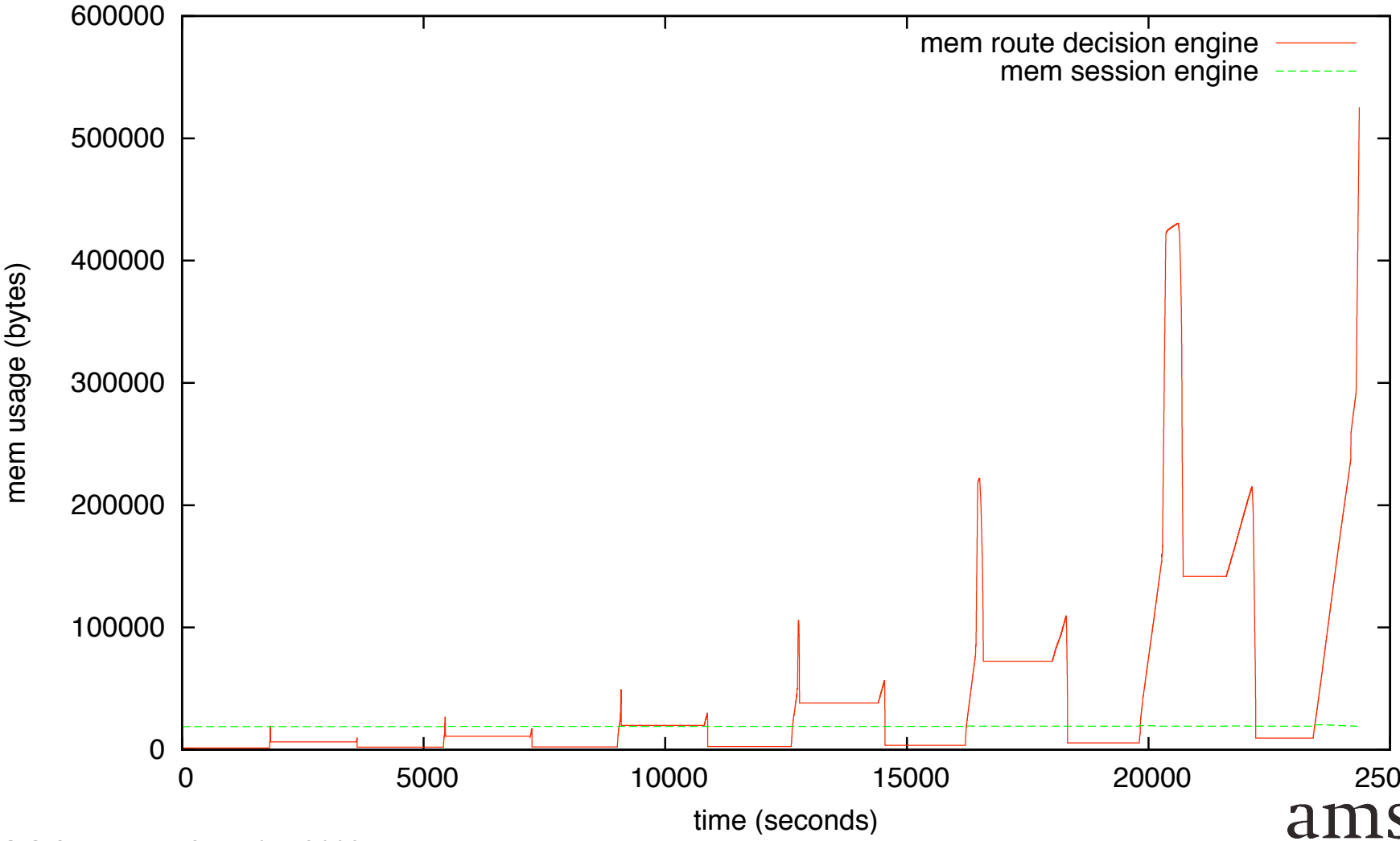
Test #4 - CPU

OpenBGPd cpu usage
multiple-rib; 253 sessions established at t=0
sent 1/2/4/8/16/32/64 updates per session beginning t=600s at a 600s interval
announce/withdraw 1 prefix each
Intel(R) Pentium(R) III CPU family 1133MHz (GenuineIntel 686-class)



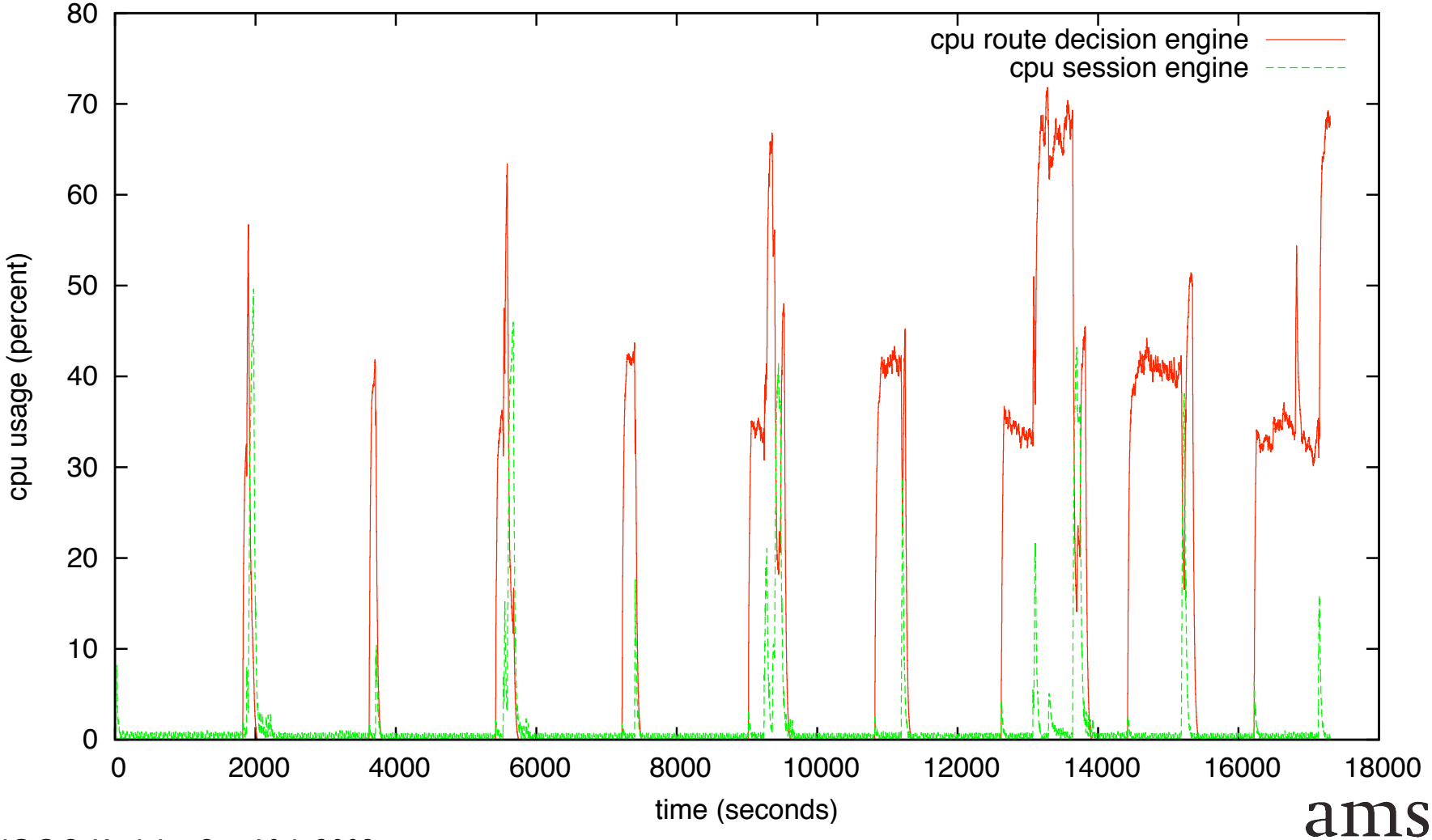
Test #4 - Mem

OpenBGPd mem usage
multiple-rib; 253 sessions established at t=0
sent 1/2/4/8/16/32/64 updates per session beginning t=600s at a 600s interval
announce/withdraw 1 prefix each
Intel(R) Pentium(R) III CPU family 1133MHz (GenuineIntel 686-class)



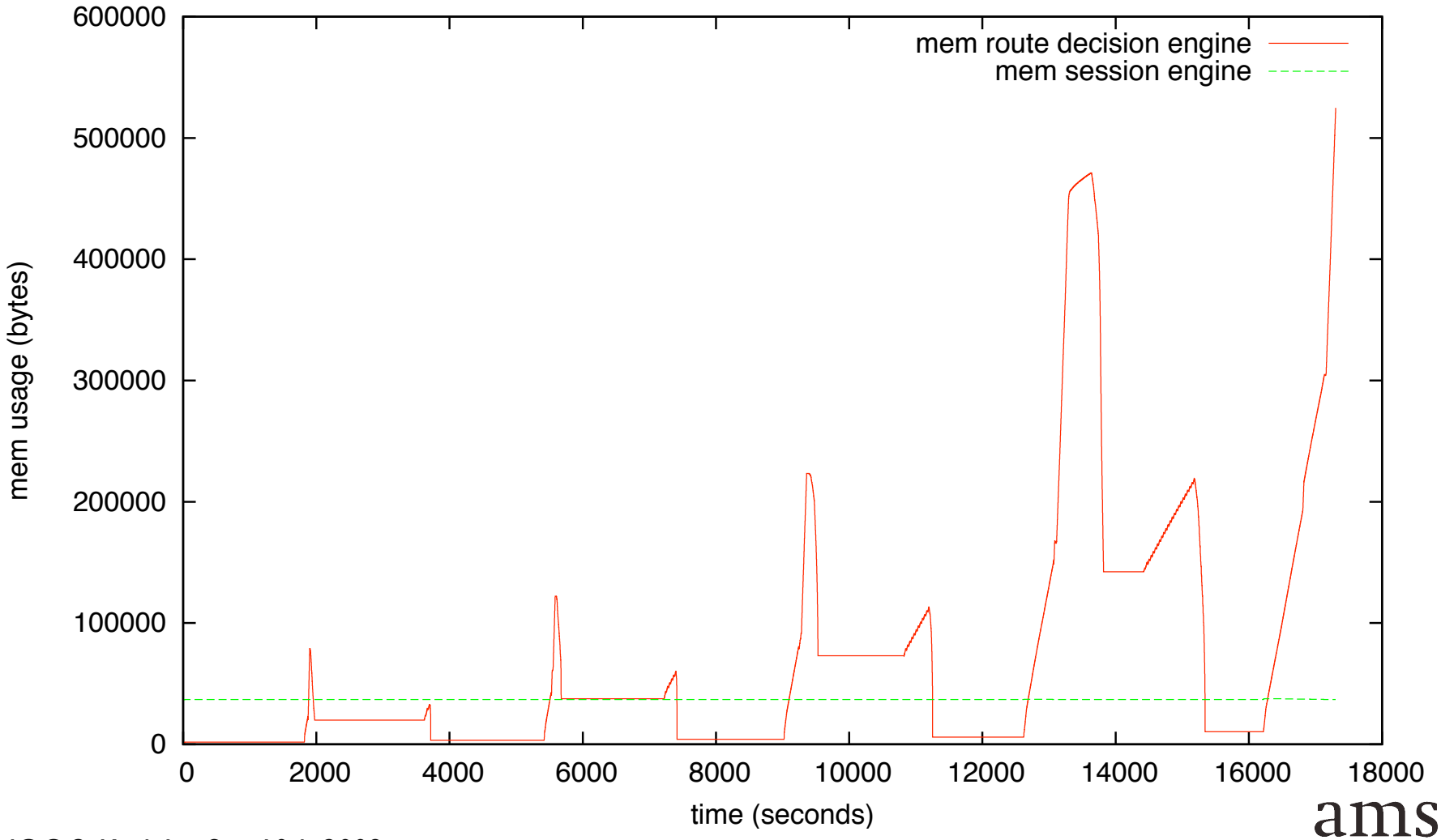
Test #5 - CPU

OpenBGPd cpu usage
multiple-rib; 506 sessions established at t=0
1/2/4/8/16/32/64 updates per sess every 30min
announce/withdraw 1 unique prefix each
Intel(R) Pentium(R) III CPU family 1133MHz (GenuineIntel 686-class)



Test #5 - Mem

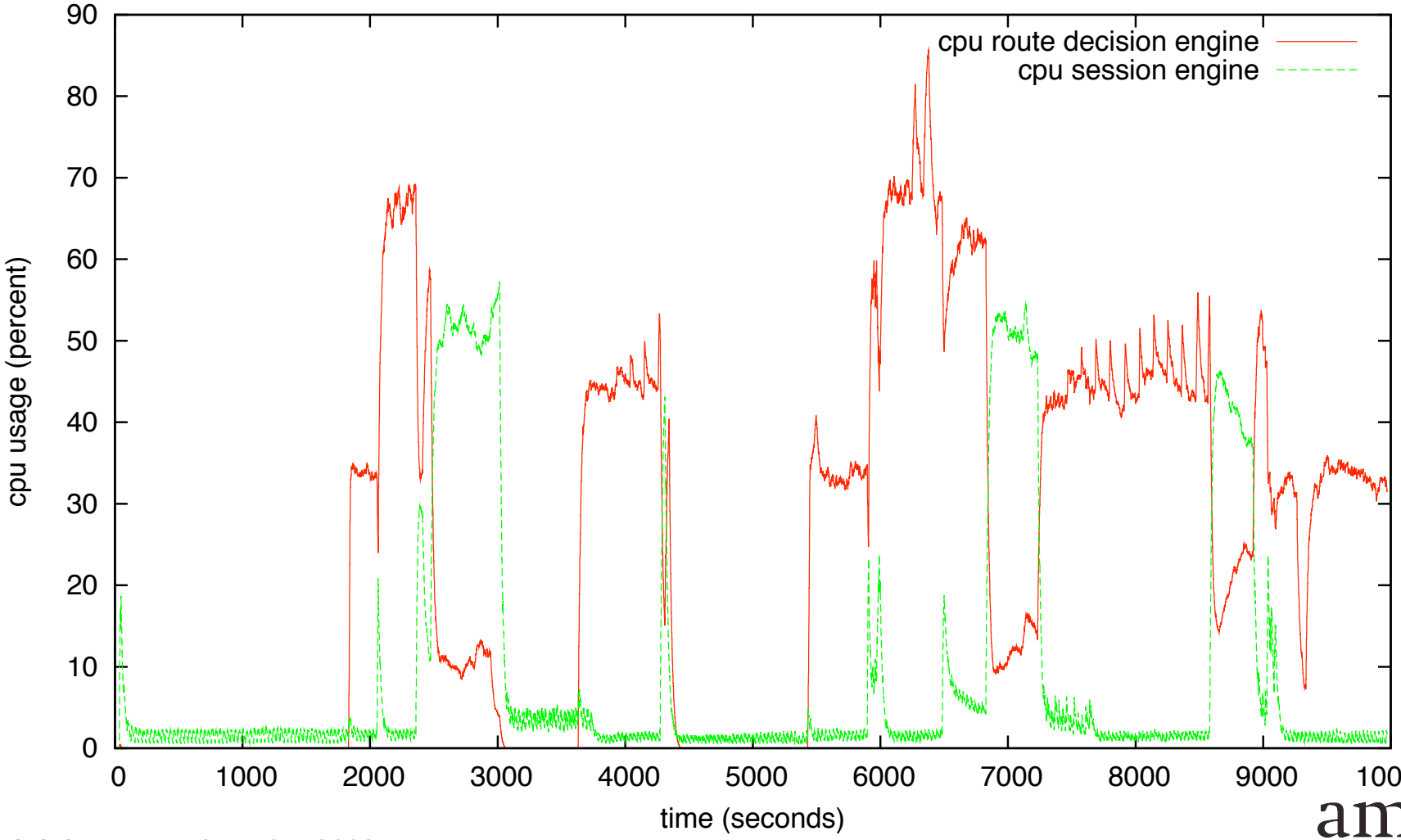
OpenBGPd mem usage
multiple-rib; 506 sessions established at t=0
1/2/4/8/16/32/64 updates per sess every 30min
announce/withdraw 1 unique prefix each
Intel(R) Pentium(R) III CPU family 1133MHz (GenuineIntel 686-class)



amsix
amsterdam internet exchange

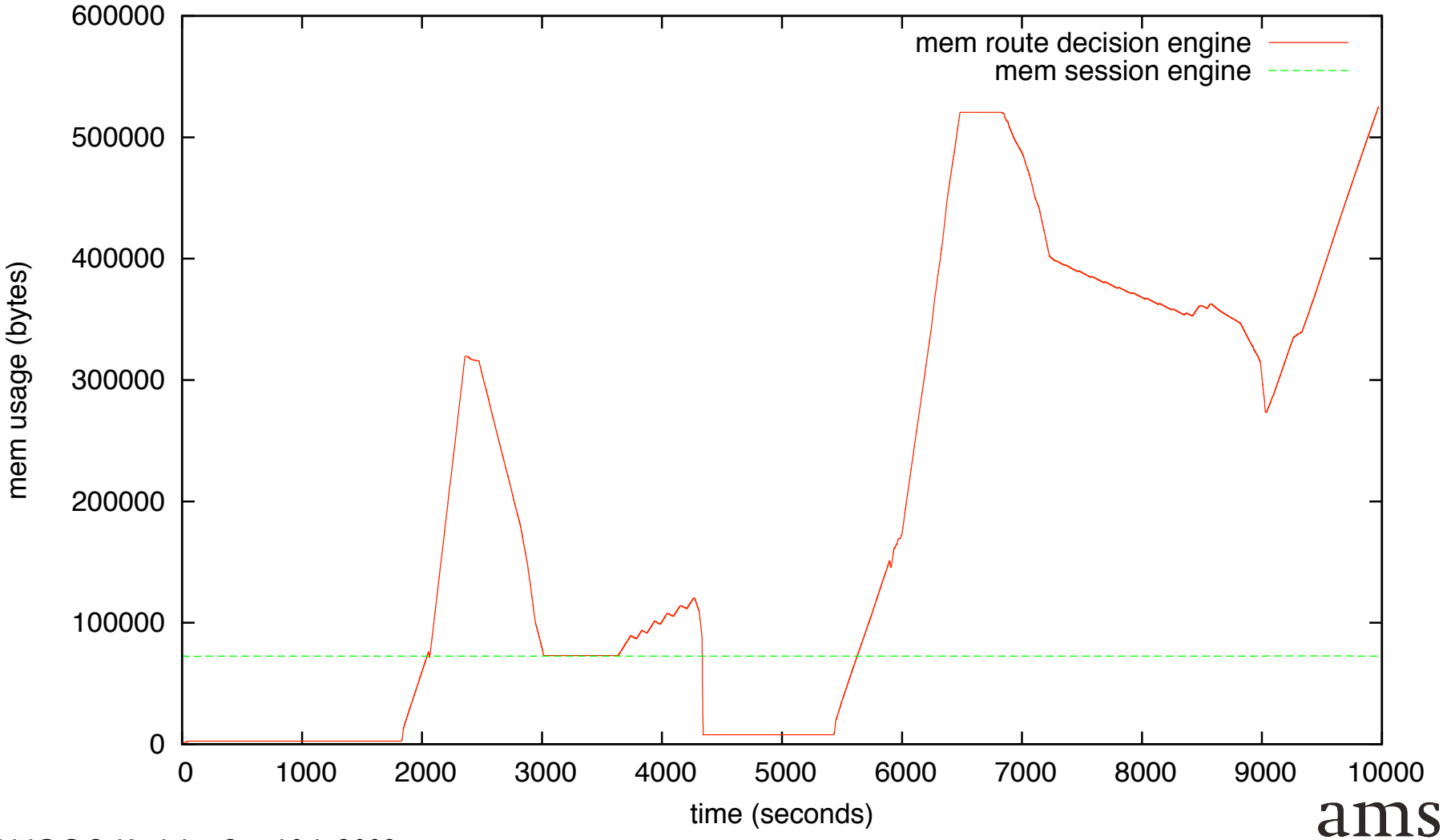
Test #6 - CPU

OpenBGPd cpu usage
multiple-rib; 1012 sessions established at t=0
sent 1/2/4/8/16/32/64 updates per session beginning t=1800s at a 1800s interval
announce/withdraw 1 prefix each
Intel(R) Pentium(R) III CPU family 1133MHz (GenuineIntel 686-class)



Test #6 - Mem

OpenBGPd mem usage
multiple-rib; 1012 sessions established at t=0
sent 1/2/4/8/16/32/64 updates per session beginning t=1800s at a 1800s interval
announce/withdraw 1 prefix each
Intel(R) Pentium(R) III CPU family 1133MHz (GenuineIntel 686-class)



Memory Usage

- Increased memory usage with the amount of UPDATEs

```
Jul  9 04:58:59 routeertnix bgpd[9200]: fatal in RDE: up_generate: Cannot allocate memory
Jul  9 04:58:59 routeertnix bgpd[7848]: Lost child: route decision engine exited
Jul  9 04:58:59 routeertnix bgpd[14688]: fatal in SE: pipe write error: Broken pipe
Jul  9 04:58:59 routeertnix bgpd[7848]: Terminating
```

- But wait... the server has 1GB memory and it crashed with 500MB!?
- Data segment size on OpenBSD: ~500MB

```
# ulimit -a
...
data seg size           (kbytes, -d) 524288
...
```

- But this is not the issue...

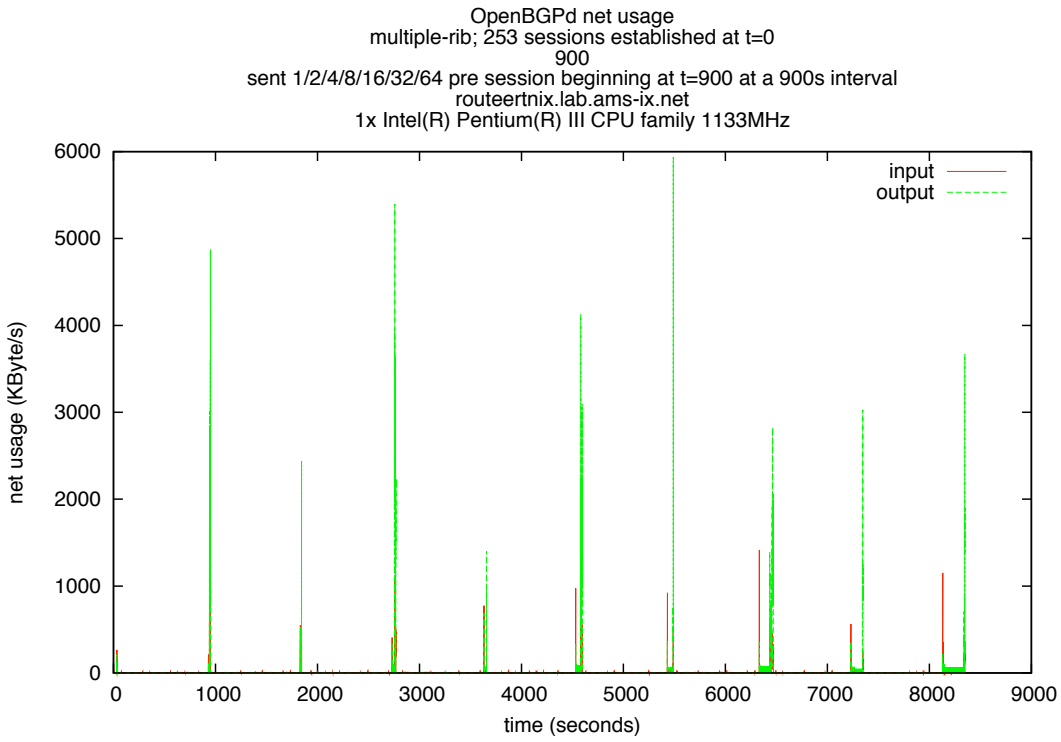
Memory Usage

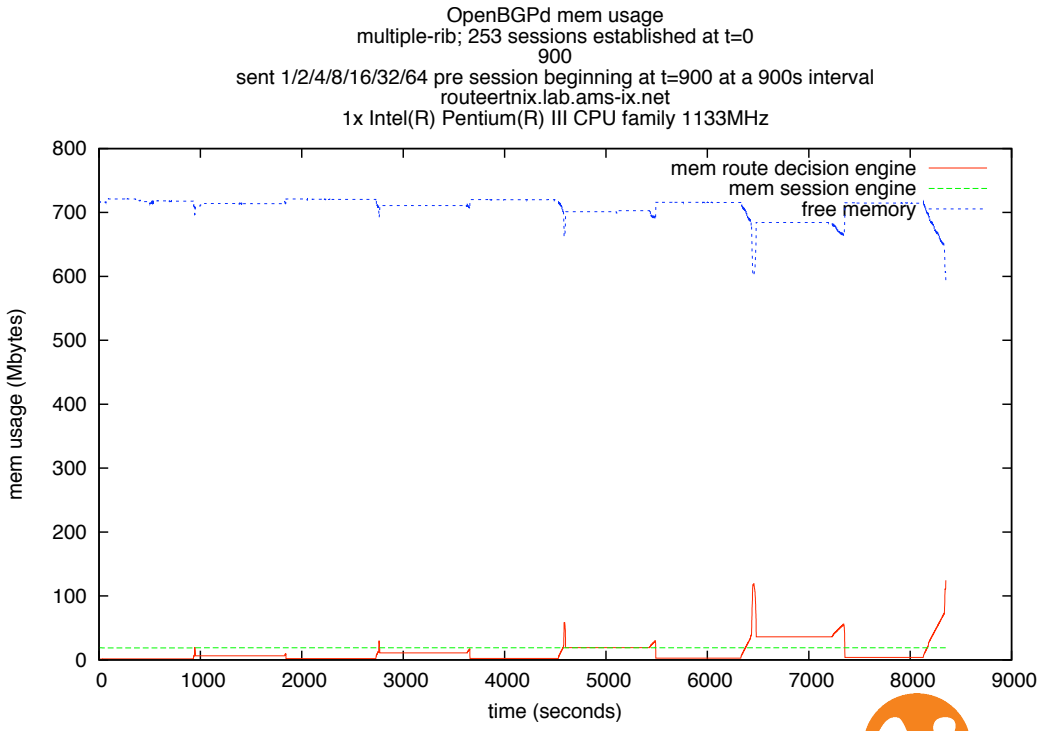
- Underlying issue seems like OpenBGPd buffers the messages before sending them
- Is the test server too slow receiving? Or OpenBSD too slow sending? Or is it the Route Server itself?
- First, lets rule out the test server, since this is what we have control over....

Memory Usage

- Perl too slow?
- Replaced socket implementation in Net::BGP with some C code
- After establishing all sessions the file descriptors are passed on to the C part
- It shows that there is not much data coming in during the time the memory occupation grows

Memory Usage vs. Net Load





PLNOG

Summary

- This is work in progress
- Many issues are still subject to investigation
- Results are not conclusive yet
- More research and updates will follow...

amsix
amsterdam internet exchange

PLNOG

Thank you!

Questions?

<elzbieta.jasinska@ams-ix.net>

amsix
amsterdam internet exchange