Running BIND or ? in large scale

Måns Nilsson,

Royal Institute of Technology, Stockholm

mansaxel@sunet.se

28 januari 2003

History

- `sunic.sunet.se` is a well-known name server, serving something for almost everybody.
- Originally the central "services" machine for SUNET, the Swedish research network.
  Of this only SMTP and DNS remains.
- Today just above 500 zones, 80% slaves, from all over the world.

Hardware & software

Several iterations have existed:

- Sun 3/50, SunOS
- Sun 4/280, SunOS
- Sun SS390, SunOS
- Sun SS690, SunOS
- Sun SS20, SunOS (here we left BIND 4)
- dig, eh, Comp, eh, HP Alpha DS20, Tru64
- We've been running BIND only, now at 8.2.7.

Thanks to Lars-Johan Liman for this list!

Recent changes

- Recursion disabled!
  – My Internet is b0rken!
  Led to lots of questions from operators too.
  You can't be too careful finding top recursion clients.
  But! Stability and safety improves; some recent security issues were irrelevant.
- BIND 9 starts to look interesting, for feature and security reasons.

(detour) Do this on your auth boxes!

```
options {
        recursion no;
};
```

...But not until you've understood the implications...

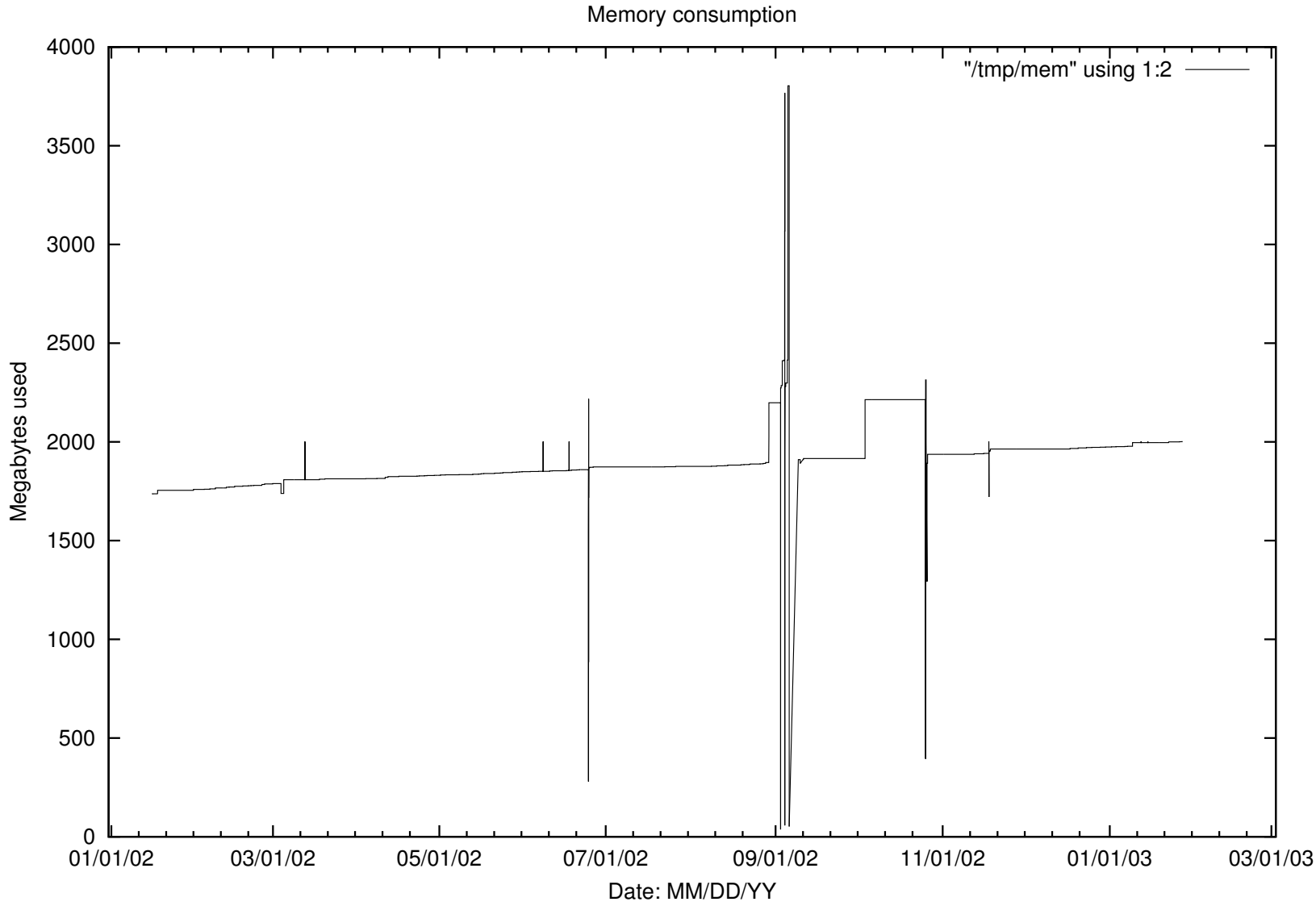
Growth

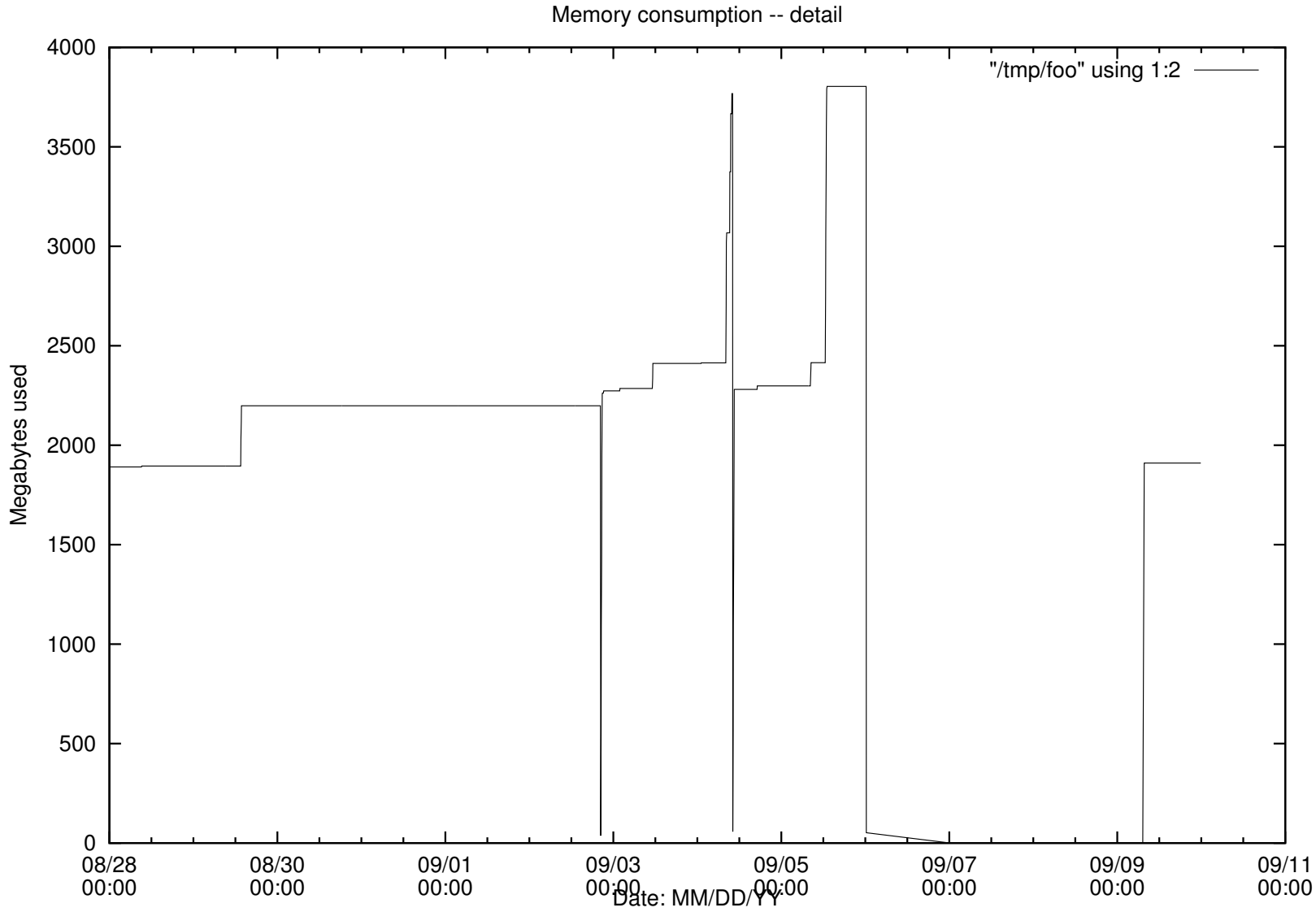
Several growth factors:

- New allocations to RIPE. We are happy to help.
- New domain structures – `ip6.arpa`, `e164.arpa`.
- Liberal ccTLD rules – SE coming, DE as always, likewise NL.
- New SUNET with lots of IP space and associated names.

Growth

- In 6 months (or so we've said, every year since 1998) DNSSec will be in production. Projected memory usage increase between 60 - 100%.
- The initially proposed software upgrade is BIND 9, which increases memory usage, and also requires transient RAM for reloads.


Memory consumption
"/tmp/mem" using 1:2
Megabytes used
4000
3500
3000
2500
2000
1500
1000
500
0
01/01/02
03/01/02
05/01/02
07/01/02
09/01/02
11/01/02
01/01/03
03/01/03
Date: MM/DD/YY


Memory consumption -- detail
"/tmp/foo" using 1:2
Megabytes used
4000
3500
3000
2500
2000
1500
1000
500
0
08/28 00:00
08/30 00:00
09/01 00:00
09/03 00:00
09/05 00:00
09/07 00:00
09/09 00:00
09/11 00:00
Date: MM/DD/YY

So, what to do?

- The hardware is still impressive. This machine is Fast!
- The memory is maxed out, and at times runs out when BIND 9 is grumpy.
- The logical path would be a new Alpha.
- But the Alpha is a dead end. (boo, hiss!)
- And, BIND 9.2.* does have `malloc()` and `free()` issues on Tru64.
- This – together – is what started the work I'm describing.

Alternative hardware

- 64bit architecture required.
- At least 8GB memory, with room for more.
- Manageability for present staff.
- Software availability and compatibility.

Alternative hardware

Leaves us with:

- AIX on RS/6000 – proven gTLD server.
- Solaris on UltraSparc – "what everybody are using."
- Tru64 has dull future.
- HP-UX / PA-RISC: Social resistance.

Software

We have a number of requirements on the software we are to run. These include, in no particular order:

- Stability.
- Compatibility; both with the specs and the Internet at large, and our management model.
- Responsive developers.
- Capacity for large data sets.
- From a purity point of view, an auth-only server is preferable.
- Price. Free is good.

Software

So, the candidates are:

- BIND 9 – "what everybody else are running, soon". Definitely interesting competitor.
- BIND 8 – Fast, but non-compliant.
- NSD – Interesting, but not there wrt large zones. Also some secondary operations issues.

Software

- djbdns – Lacks features and social interface.
- PowerDNS – Not tested, yet, but will be during spring 2003. Does promise a lot.
- Microsoft Windows 2000 DNS – Not considered. At all.
- Nominum ANS – currently under test.

ANS

ANS is a new, written-from-scratch name server.

- Auth-only – good.
- Not free – bad.
- Untested – bad.
- Responsive
  developers – good.
- Built for
  large zones – good.
- Shows good
  performance – good.

- Slow to start up – bad
- Requires not-yet-written support system to support BIND-targeted large zones – bad.
- Moves performance bottleneck from RAM size to disk i/o. – might be good or bad.

Conclusions

- So, are there any?
- No, not really. We are still considering our alternatives. It takes time...
- Nominum is helping us (and them) to make ANS work.
- We will have to buy some new hardware, regardless. This we know.

The End

Questions?

mansaxel@sunet.se