

Linux Kongress 2008

Hamburg/Germany

Robert Olsson Uppsala University

2008-10-10

Over 10 years in production

- Three major installations
- UU core routers towards SUNET
- UU Student Network 30.000 students
- [ftp.sunet.se](http://sunet.se)

Over 10 years in production

UU facts

Over 25.000 registered hosts

Dual ISP BGP connect GIGE

Local DMZ BGP peering GIGE

Ipv4/Ipv6

OSPFv2/OSPFv3

600 netfilter rules

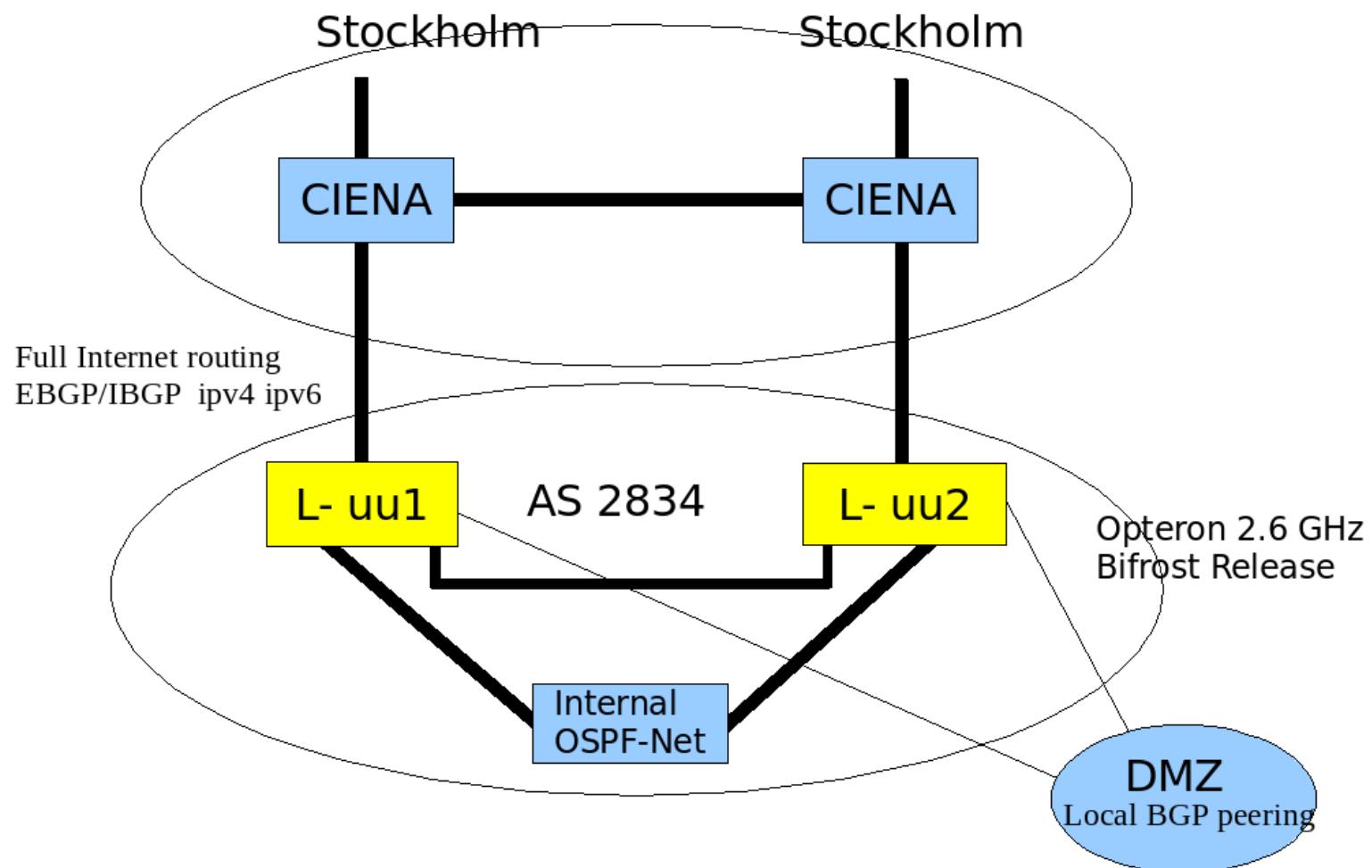
10 Cisco 6500 OSPF-routers

Redundant Power

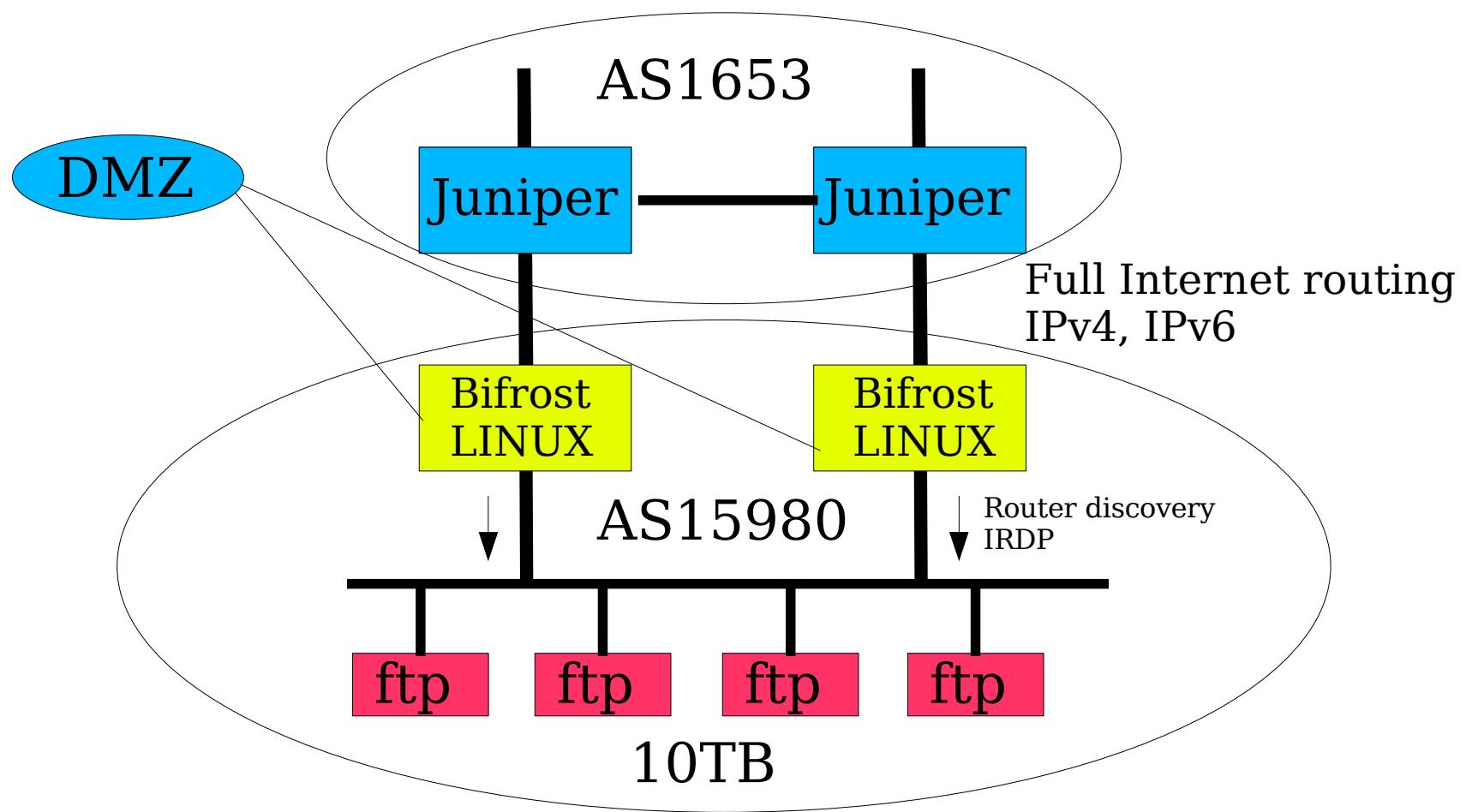
10g planned

Over 10 years in production

BGP topology at Uppsala University



Over 10 years in production The SUNET FTP ARCHIVE



Over 10 years in production

Student Network facts

Dual ISP BGP connect GIGE

Local DMZ BGP peering GIGE

Ipv4

IRDP (ICMP)

About 30 netfilter rules

19 netlogin-service boxes for premises

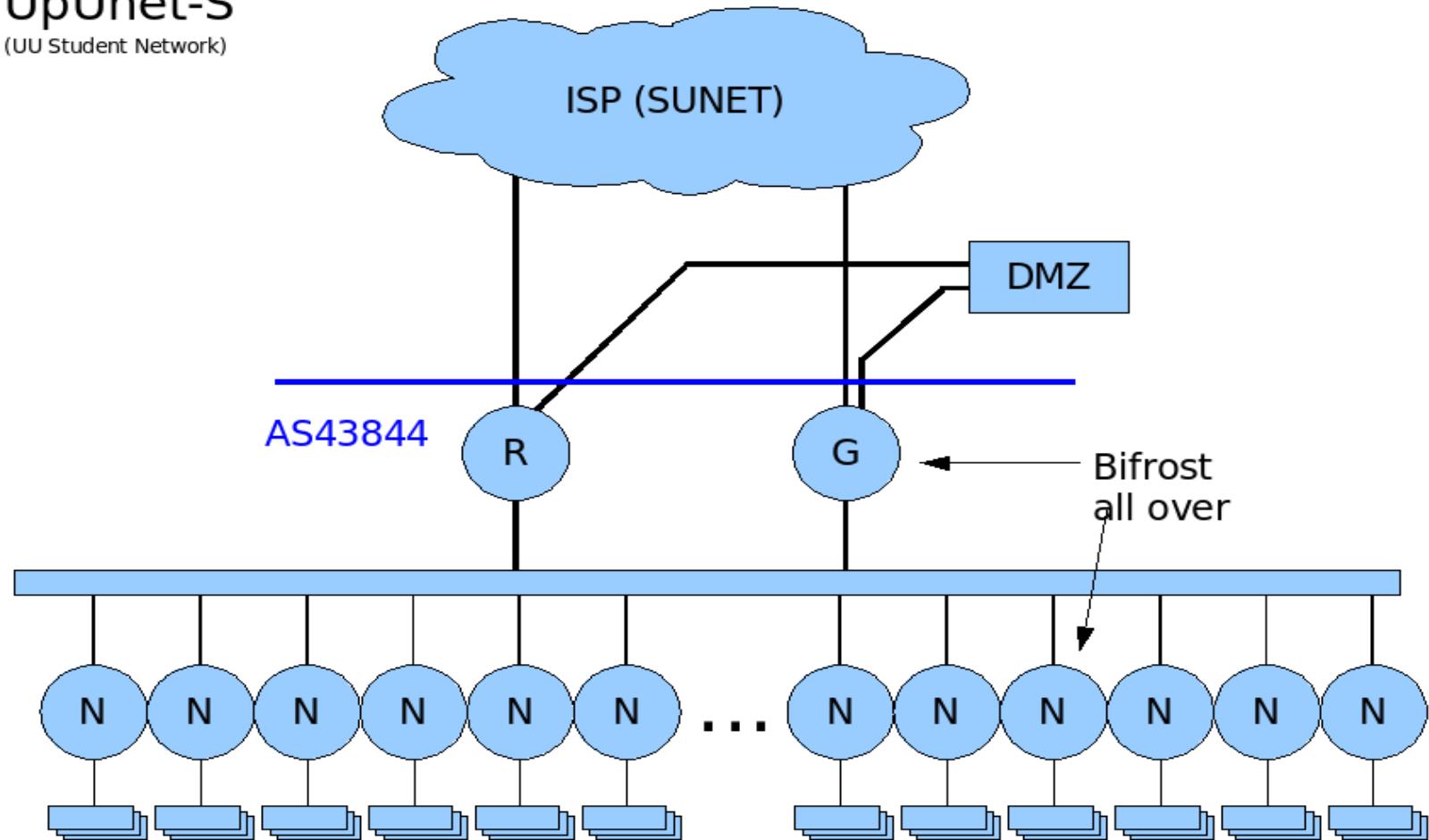
Very “innovative” users

Well connected

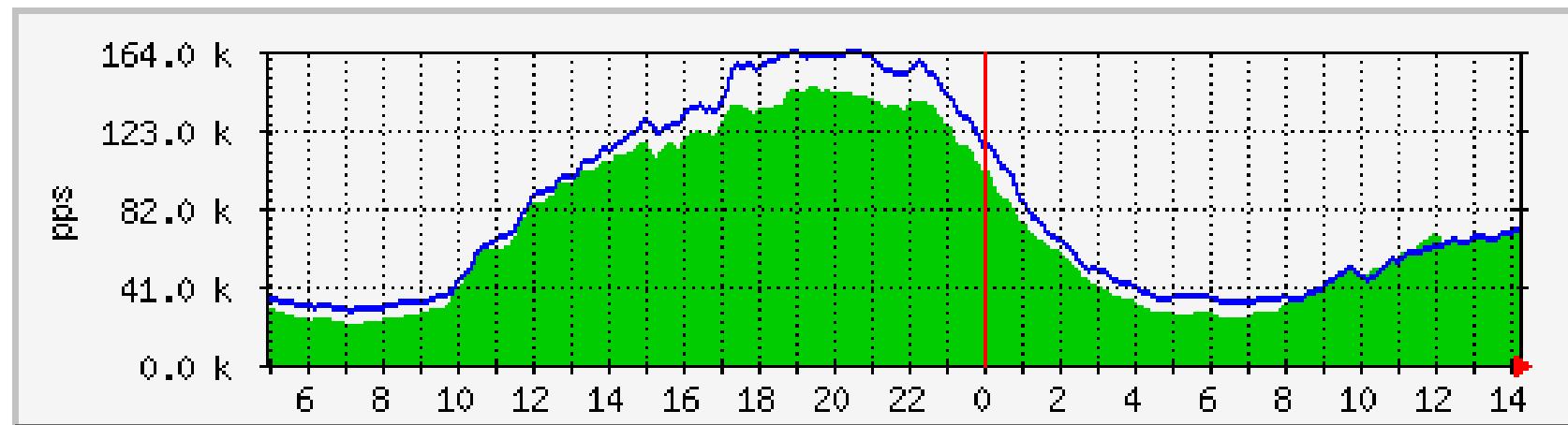
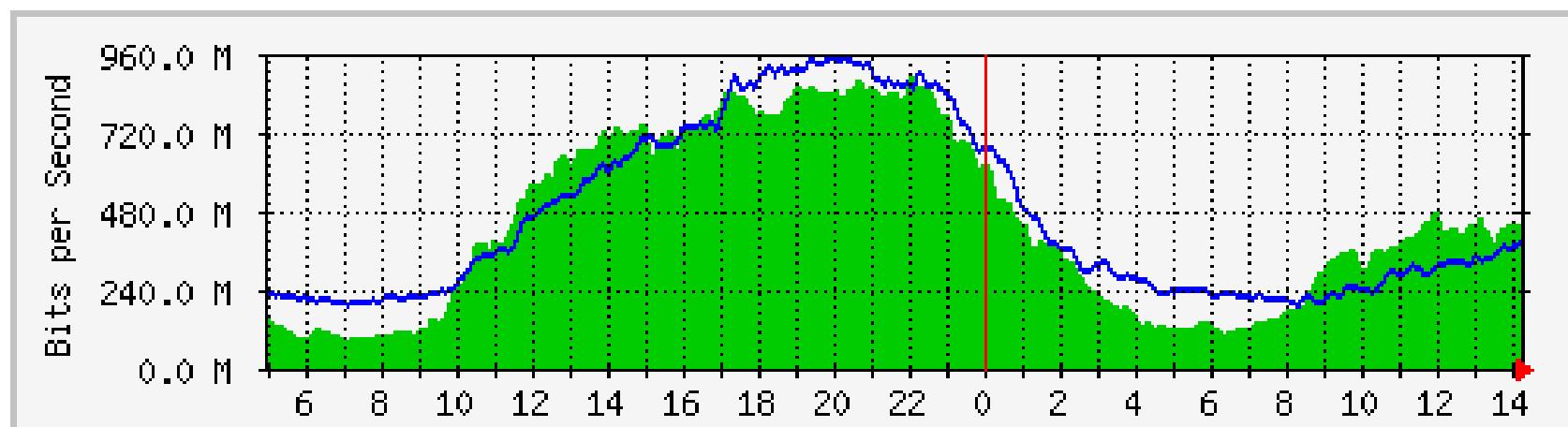
10g planned

Over 10 years in production

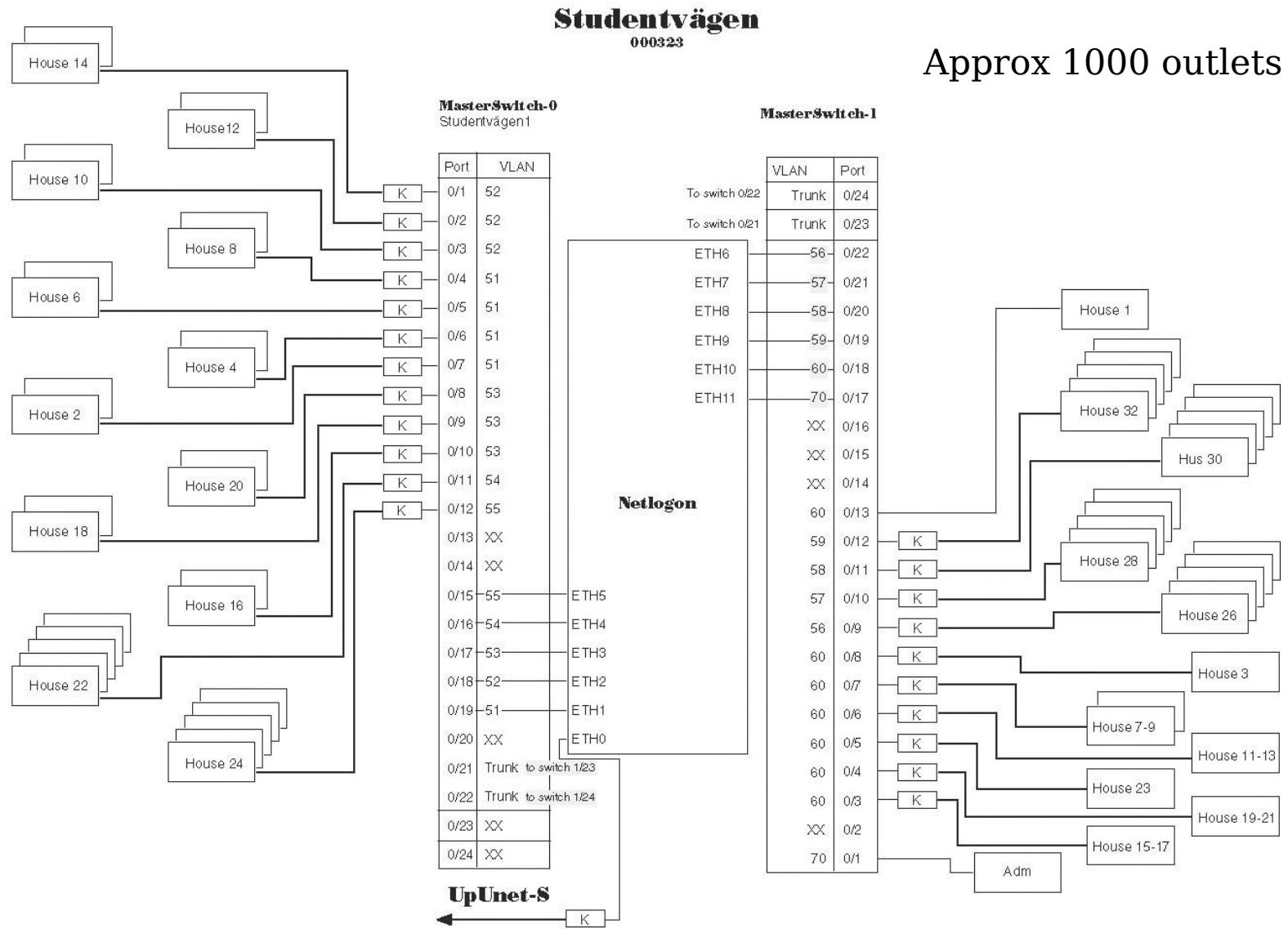
UpUnet-S
(UU Student Network)



Over 10 years in production Student Network Core Router



IP-login installation at Uppsala University



Testing, Verification Development & Research

- Started out as simple testing.
- Curiosity, Open Source, Collaboration
- Relatively freedom, the idea to use in own infrastructure. No need for external funding.
- OS was intended for desktops.

Building Blocks

Hardware:

PC

- Motherbord/CPU/Memory
- Network Interfaces
- GIGE/10g WiFi etc

Software

- Operating System
- Linux/BSD/Microsoft
- Applications
- Routing Daemons
- Quagga/XORP
- IP-login/netlogon

Network

- Cable, Fiber, Copper
- Equipment, Switches

Testing, Verification Development & Research

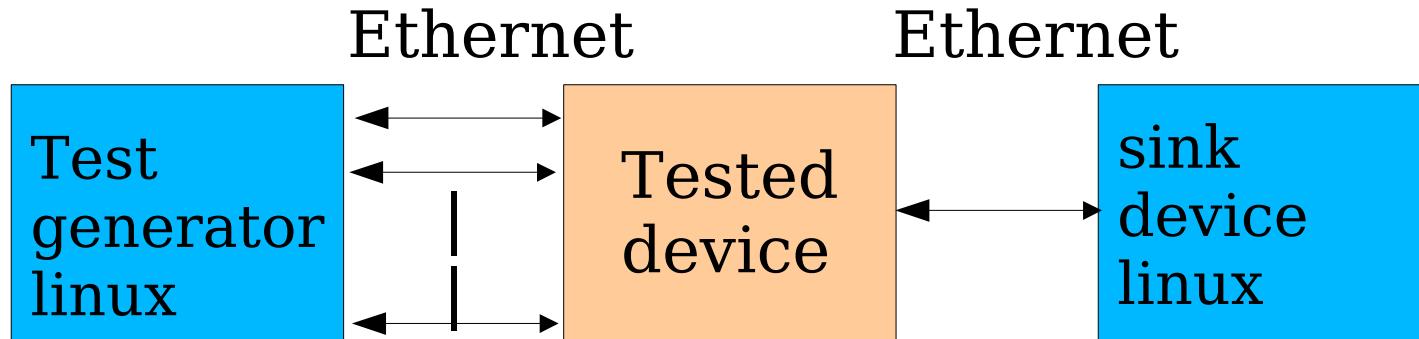
No need for test network. We could test in own infrastructure. (Or SLU)

We could work on complicated issues

- NAPI 3 years
- Pktgen 2 years
- fib_trie 1 year
- TRASH 1 year
- Hardware Testing Many years

Flexible netlab at Uppsala University

El cheapo-- High customizable -- We write code :-)

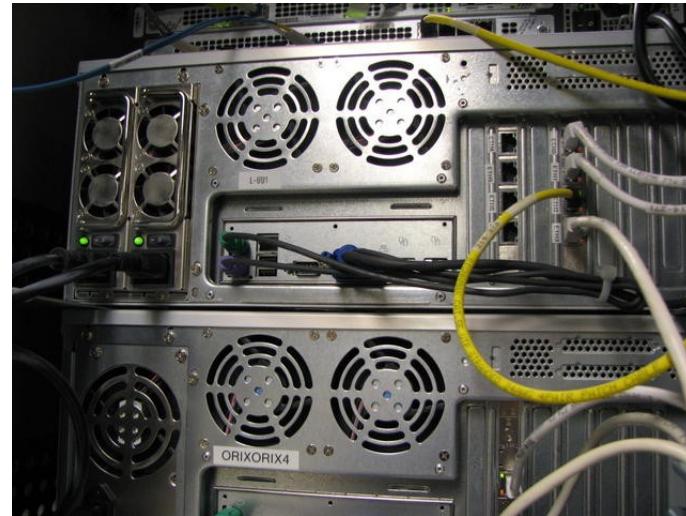


- * Raw packet performance
- * TCP
- * Timing
- * Variants

netlab at UU



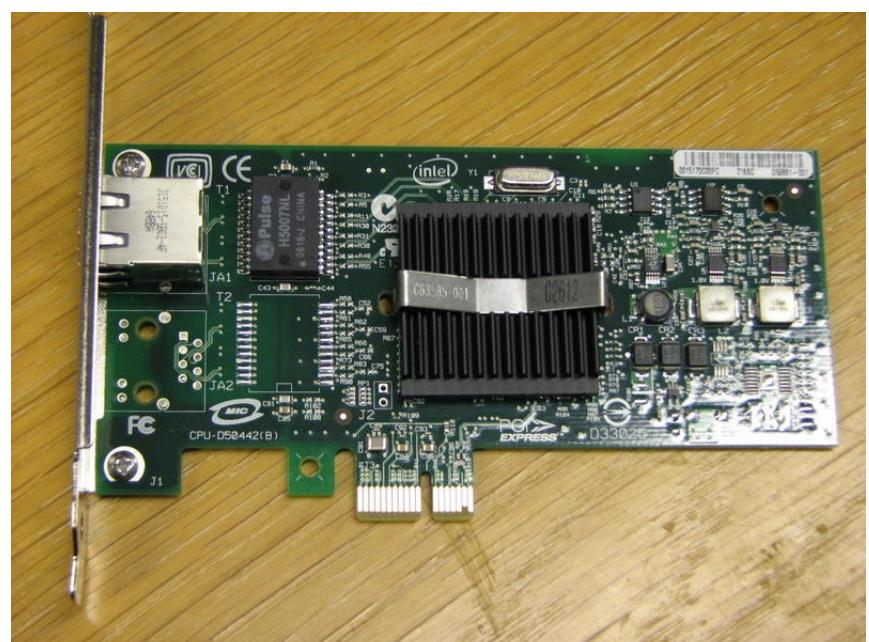
Dual-Power supply



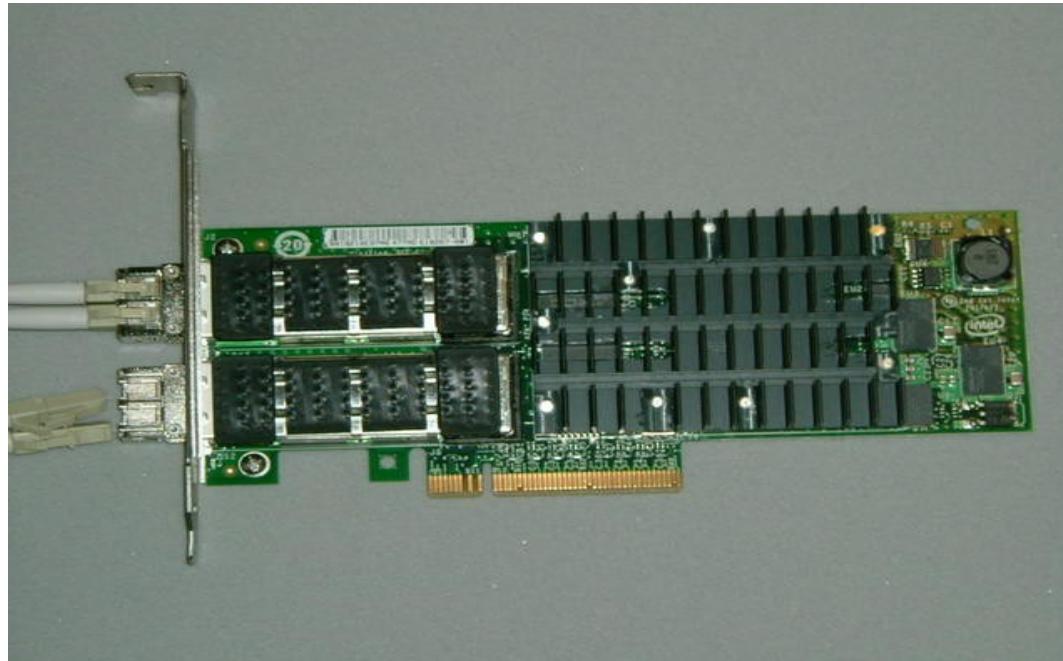
PIII for many years
[ftp.sunet.se](ftp://ftp.sunet.se)



Intel NIC's



Latest & Greatest Hardware



Intel 10g board Chipset 82598

Open chip specs. Thanks Intel!
But why fixed XFP's??
Better classifier needed.

Latest & Greatest Hardware



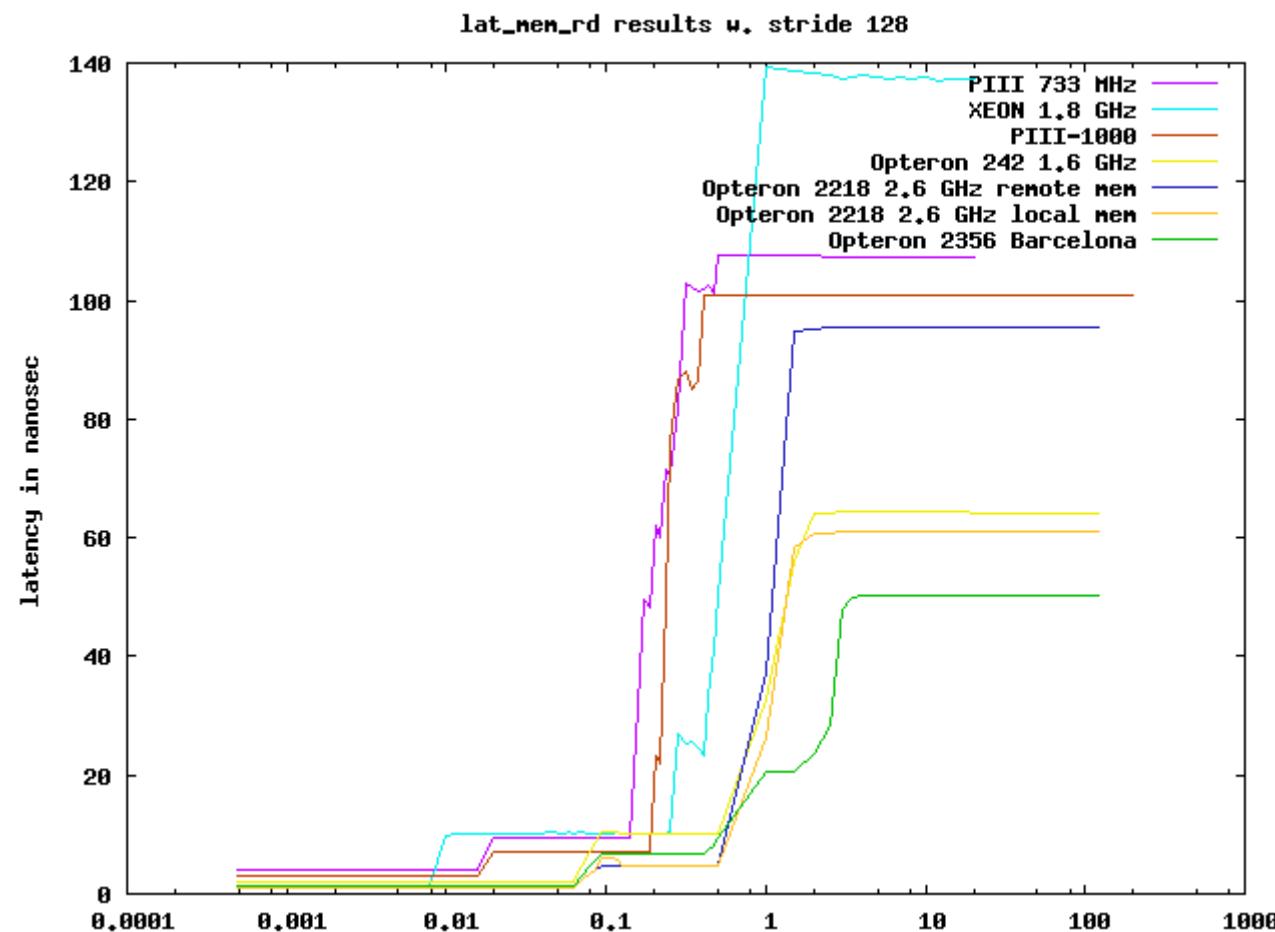
2U Hi-End Opteron box
TYAN S2927/Barcelona

Not all were blessed...

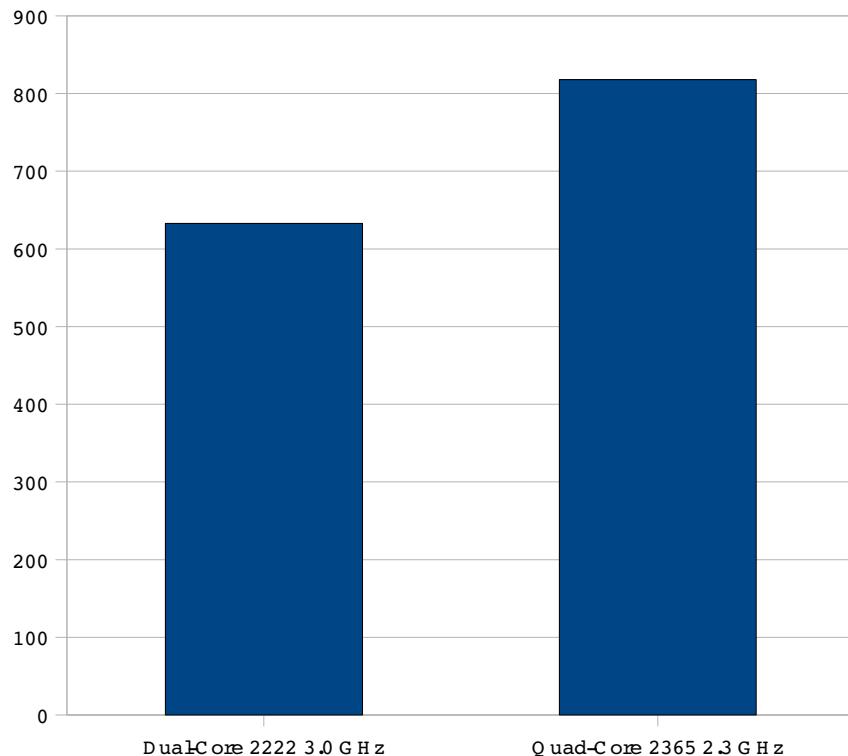


Memory Latency

lat_mem_rd from LMbench



Quad vs Dual Core Opteron



Surprising!

One CPU core on 2.3 GHz
is faster than the 3.0 GHz
Dual-Core.

L3 cache, Microcode?

2U Hi-End Opteron box
TYAN S2927/Barcelona

Bifrost concept

- Linux kernel collaboration
- Performance testing, development of tools and testing techniques
- Hardware validation, support from big vendors
- Detect and cure problems in lab not in the network infrastructure.
- Test deploy (Often in own network)

The Linux Ashram



The guru is ANK

Kernel footprints

HW_FLOWCONTROL
Tulip

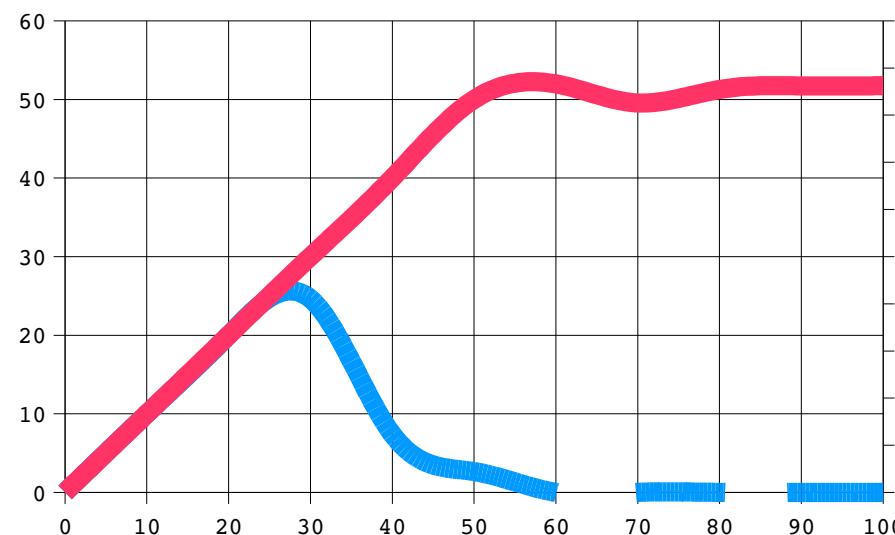
FASTROUTE path

Whitehole device. In the middle of dev.c
Hardwired IP addresses. (Russian?)

Overall Effect

- Inelegant handling of heavy net loads
 - System collapse
- Scalability affected
 - System and number of NICs
 - A single hogger netdev can bring the system to its knees and deny service to others

Summary 2.4 vs feedback

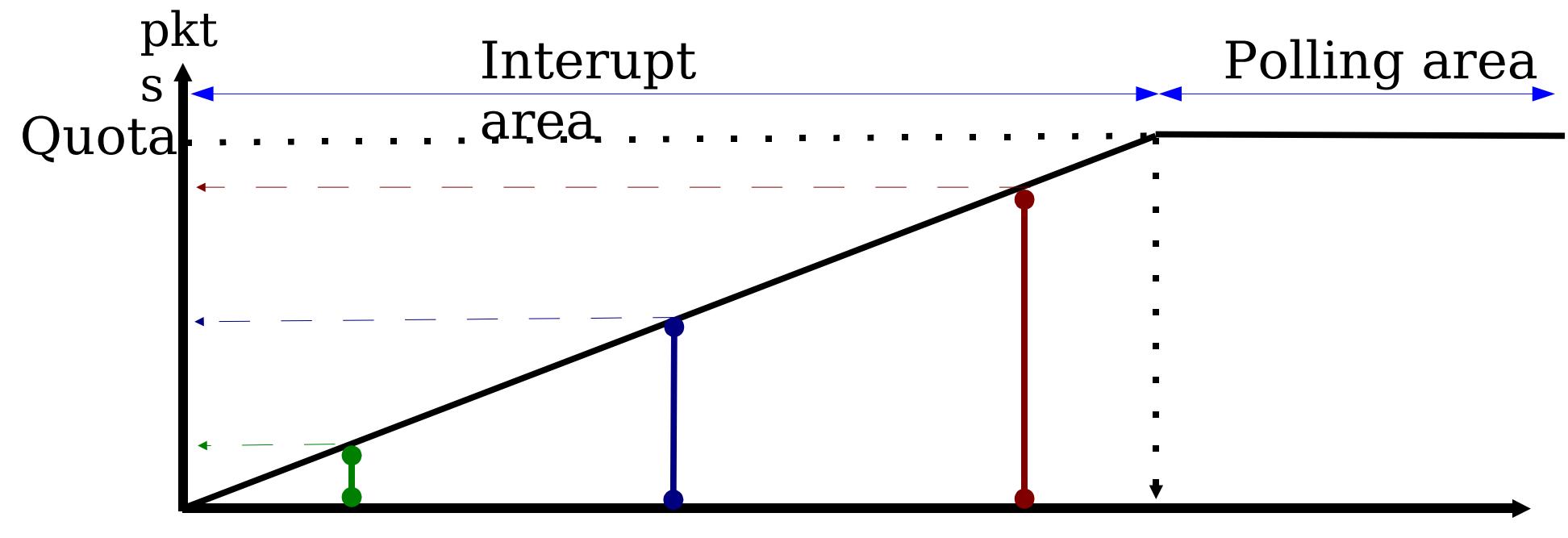


March 15 report on lkml

Thread: "How to optimize routing performance"
reported by

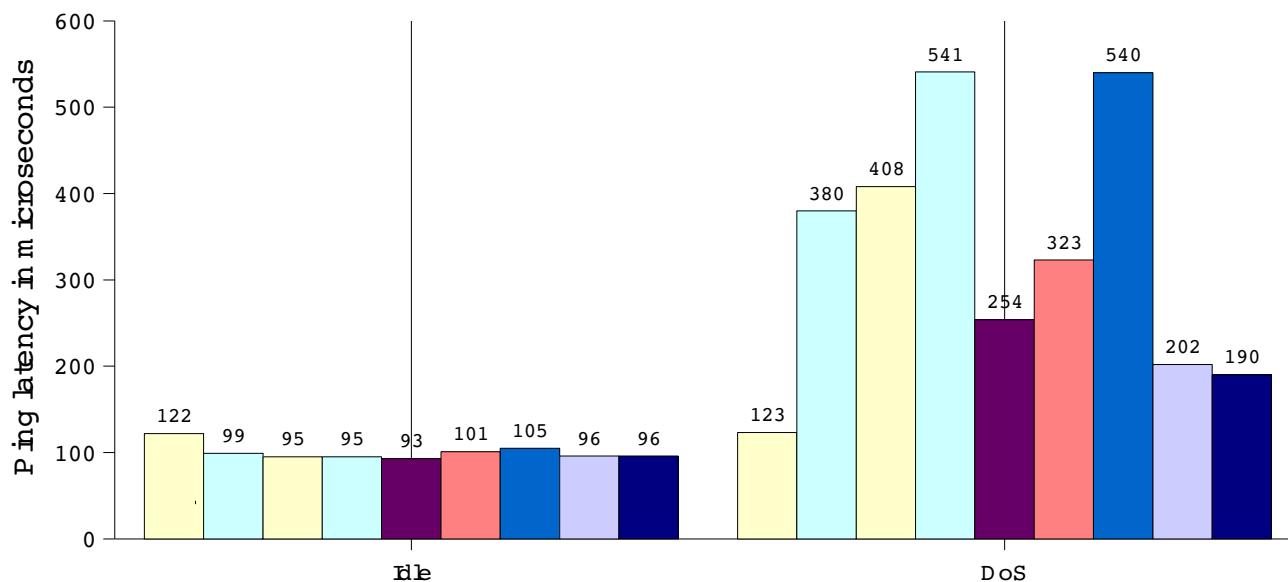
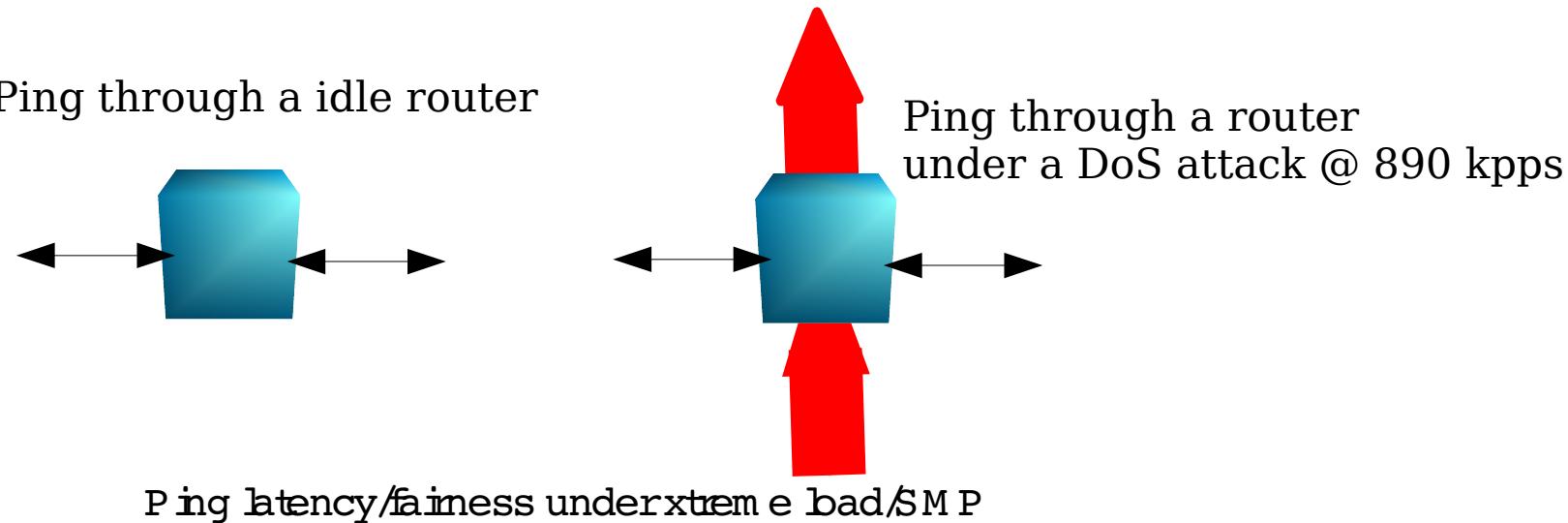
Marten.Wikstron@framsfab.se
- Linux 2.4 peaks at 27Kpps
- Pentium Pro 200, 64MB RAM

A high level view of new system



- P packets to deliver to the stack (on the RX ring)
- Horizontal line shows different netdevs with different interrupt priorities
- Area under curve shows how many packets before next interrupt
- Quota enforces fair share

NAPI observations & issue: fairness



Very well behaved just an increase a couple of 100 microsec !!

NAPI Kernel support

NAPI kernel part was included in:
2.5.7 and back ported to 2.4.20

Current driver support:

e1000 Intel GIGE NIC's – (UFO driver)

First driver where (RX & TX done in softirq)

tg3 BroadCom GIGE NIC's

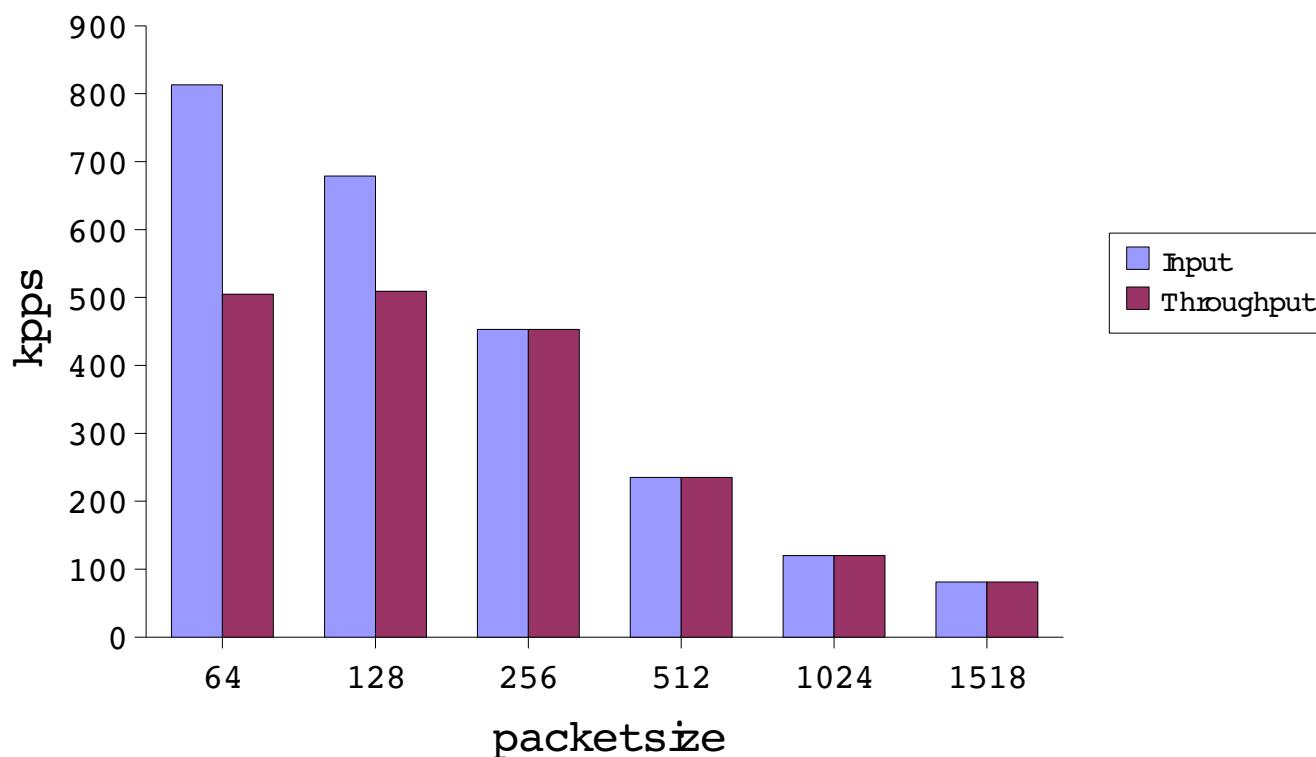
dl2k D-Link GIGE NIC's

tulip (pending) 100 Mbs

Forwarding performance (old)

Linux forwarding rate at different pkt sizes

Linux 2.5.58 UP /skb recycling 1.8 GHz XEON

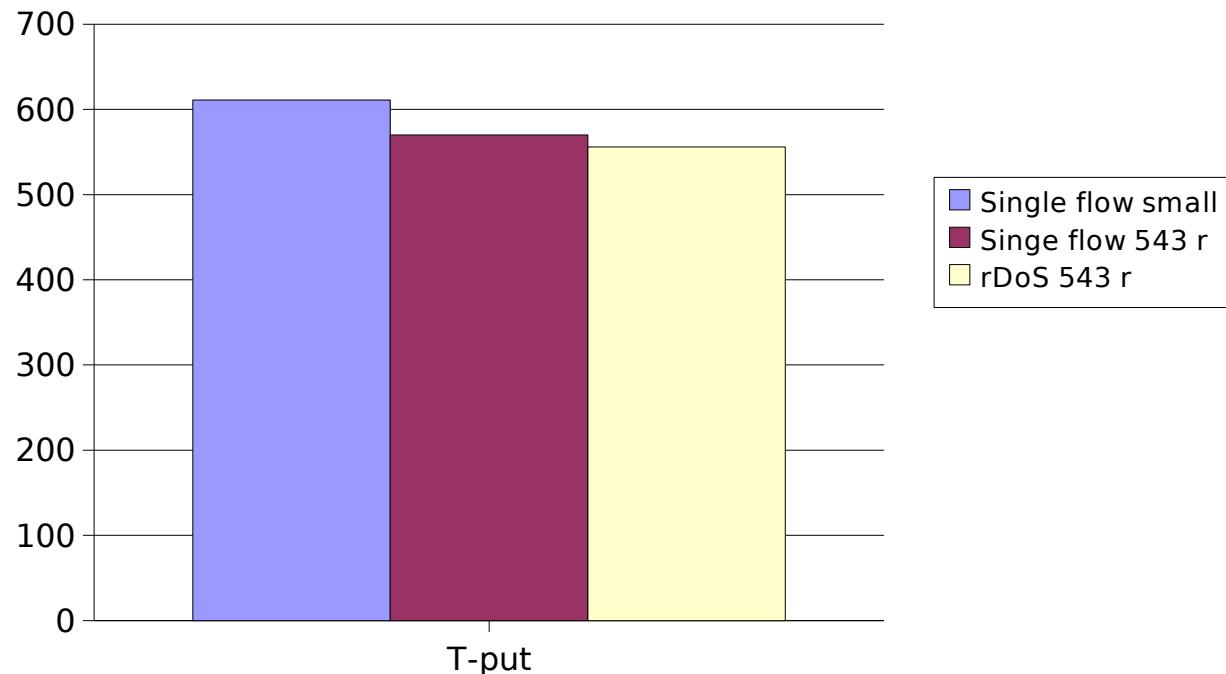


Fills a GIGE pipe -- starting from 256 byte pkts

ipv6 performance(old)

Forwarding kpps 76 byte pkt.

Linux 2.5.12 1 CPU(SMP) Opteron 1.6 GHz e1000

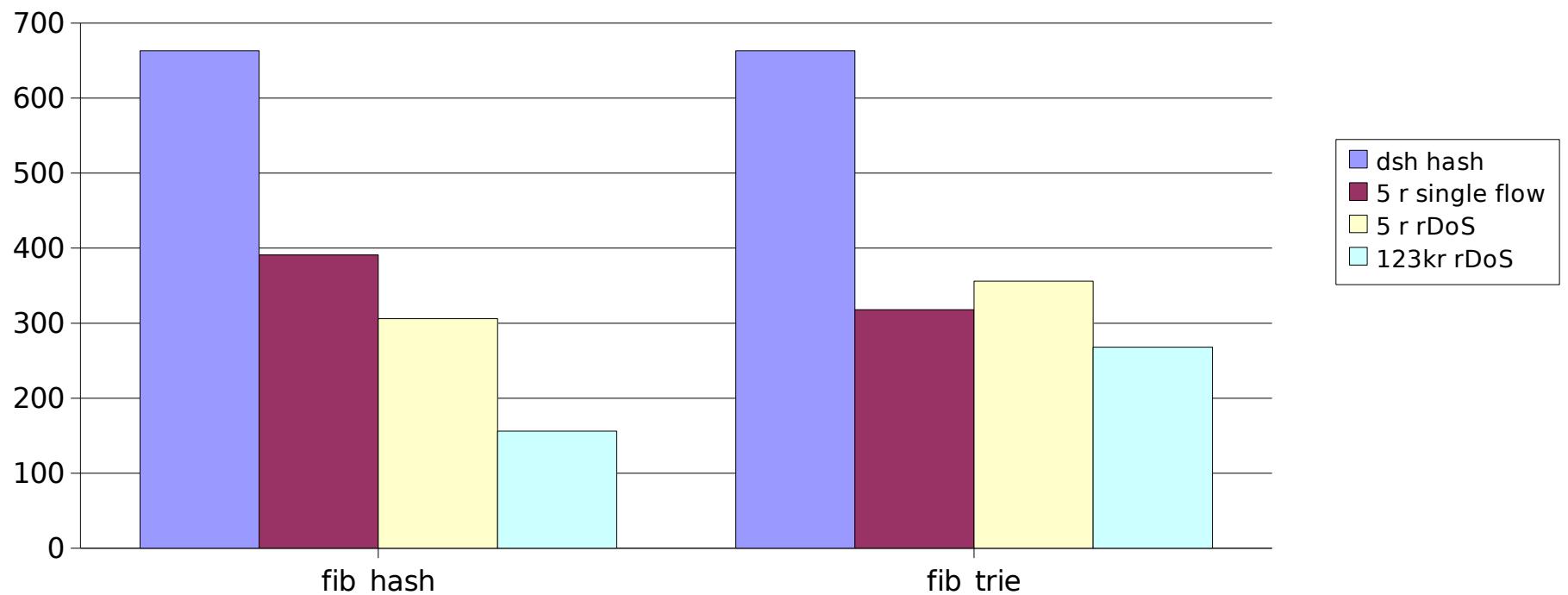


How rDoS work on sparse routing table?

fib_trie performance comparison

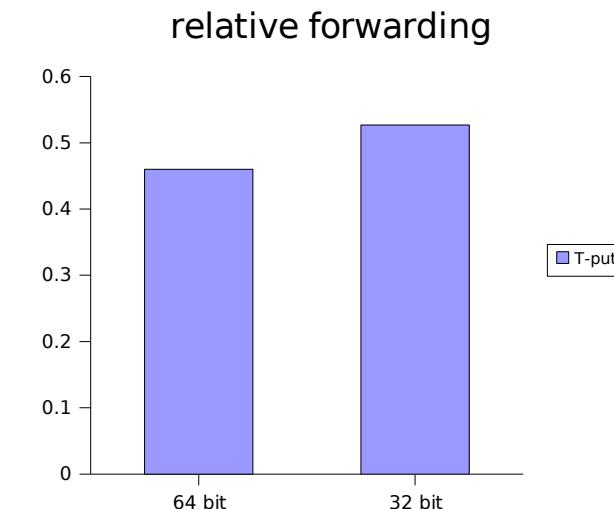
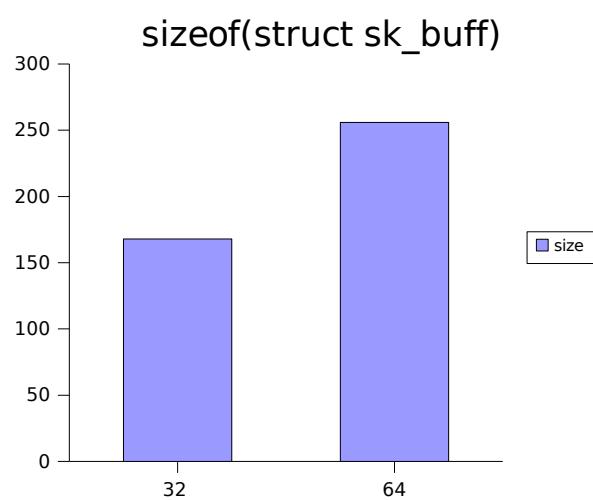
forwarding kpps

Linux 2.6.16 1 CPU used(SMP) Opteron 1.6 GHz e1000



Preroute patches to disable route hash

32/64 bit || sizeof(sk_buff)



Gcc 3.4 x86_64 vs i686 on same HW

Trash data-structure

Interesting novel approach. Trie-Hash --> Trash

When extending the LC-trie

Paper with Stefan Nilsson/KTH

Exploits that key-length does not affect tree depth

We lengthen the so key it can be better compressed.

Implemented in Linux forwarding patch as a replacement to the route hash.

Trash data-structure

Can do full key lookup. src/dst/sport/dport/proto/if
etc and later socket.

For even ip6 with little performance degradation

Could be a candidate for the grand unified lookup

Full flow lookup can understand connections.

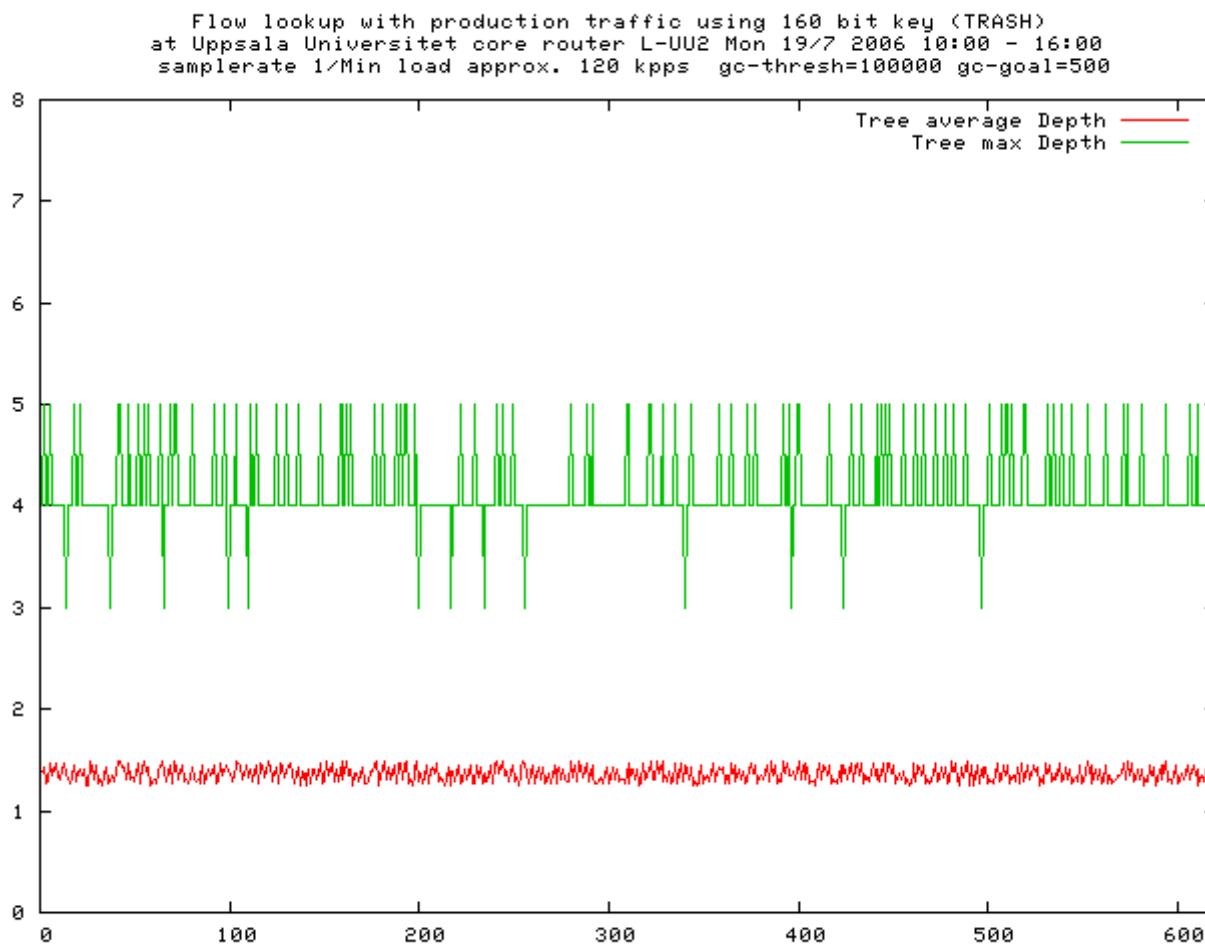
Free flow logging etc

New garbage collection (GC) possible. Active GC
stated

AGC in the paper. Listen to TCP SYN, FIN and RST
Show to be performance winner.

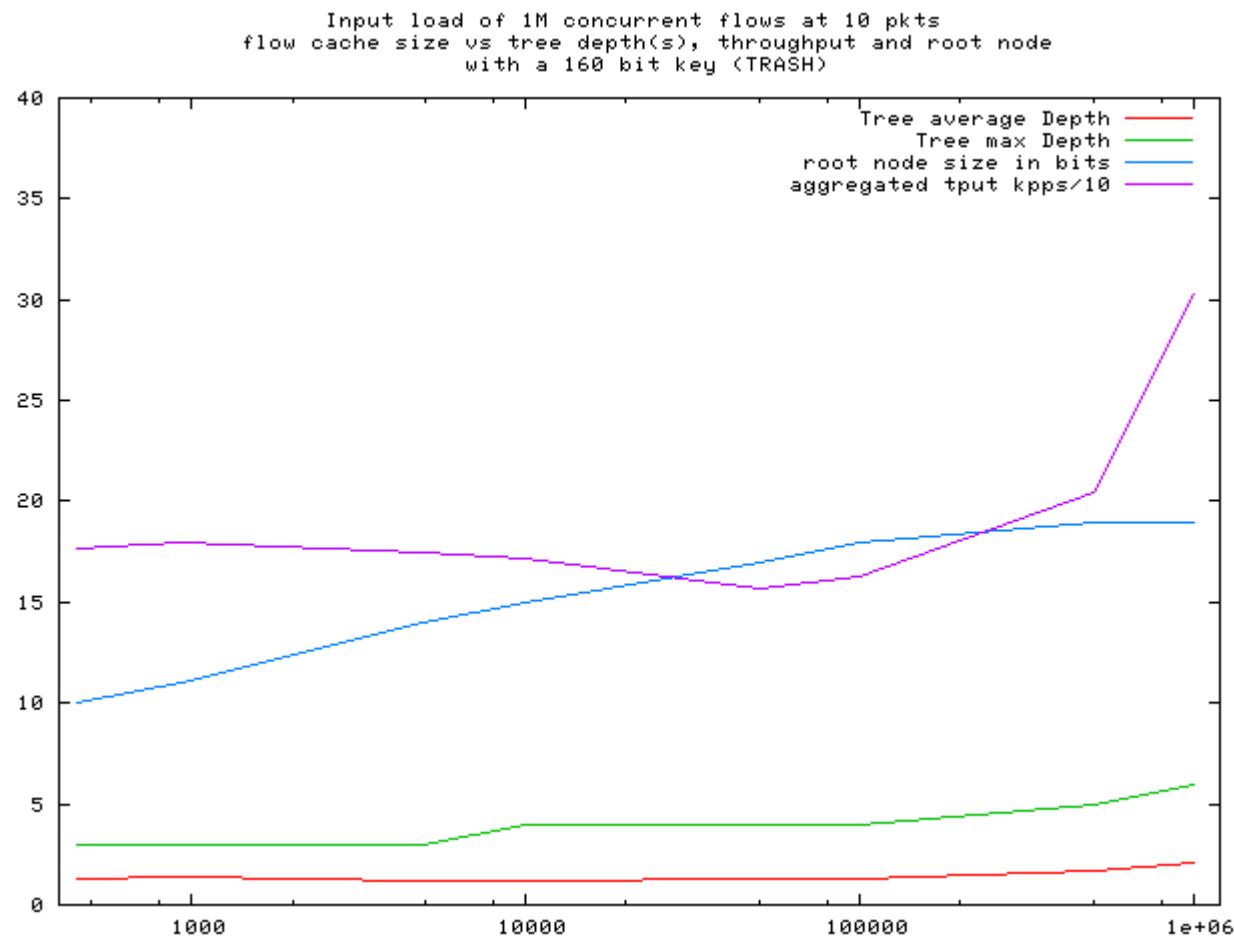
Trash data-structure

Uppsala Universitet core router



Trash data-structure

Very flat(fast) trees



Fully parallel router multi-queue breakthrough

Load from one incoming 10g interface can be split among several CPU-cores

Using RSS (Receiver Scale Option). New NIC HW classifier

MSI-X interrupts affinity for RX, TX so a packet a skb is handled by one CPU core.

Breakthrough forwarding and for networking in general.

Fully parallel router concept

multi-queue breakthrough

In experiment we used Intel 82598 adapters.
Intel follows MS NDIS 6.0 for virtualization

SUN's 10g board has a more potent HW classifier
aka TCAM.

Potent classifiers can yet another breakthrough for
both functions and performance.

Control plane separation, (routing daemons)
QoS, filters etc.

Fully parallel router multi-queue breakthrough

Flow load. 31.000 fib_lookups/sec

BGP table w. 271.064 routes

Different 3 packet sizes

64 bytes 45%

576 bytes 25%

1500 bytes 30%

RSS and Multi-Queue (RX and TX) in use

Linux 2.6.27-rc2 ixgbe-1.3.31.5 + patches

Using 2/4 CPU cores from AMD Barcelona 2.3 GHz

Forwarding:: 6.2 Gbit/s (960 kpps)

10g boards

multi-queue breakthrough

SUN's seems to use XFP's. Anyone using it....

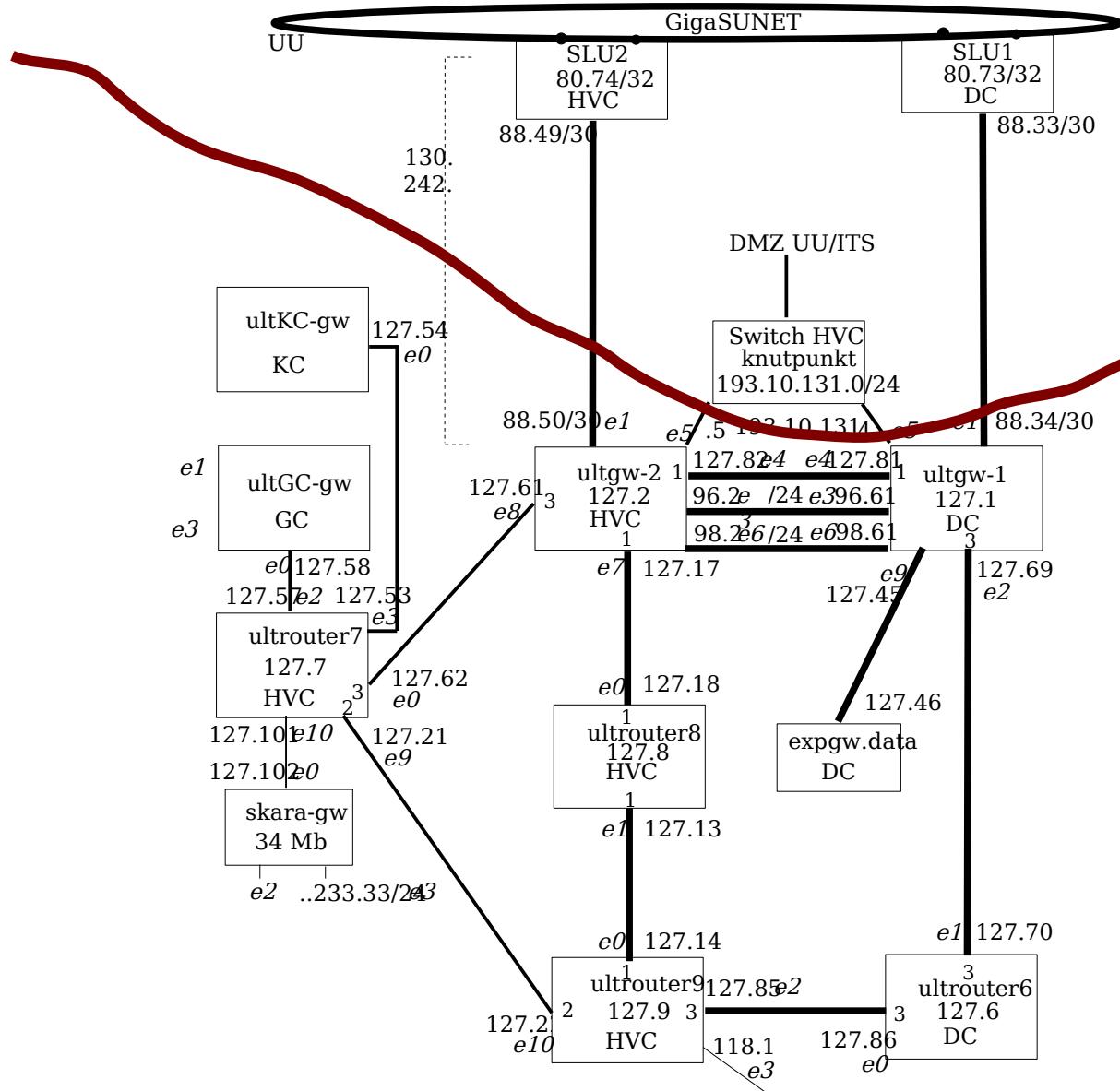
Other boards with SFP/SFP+/XFP ??

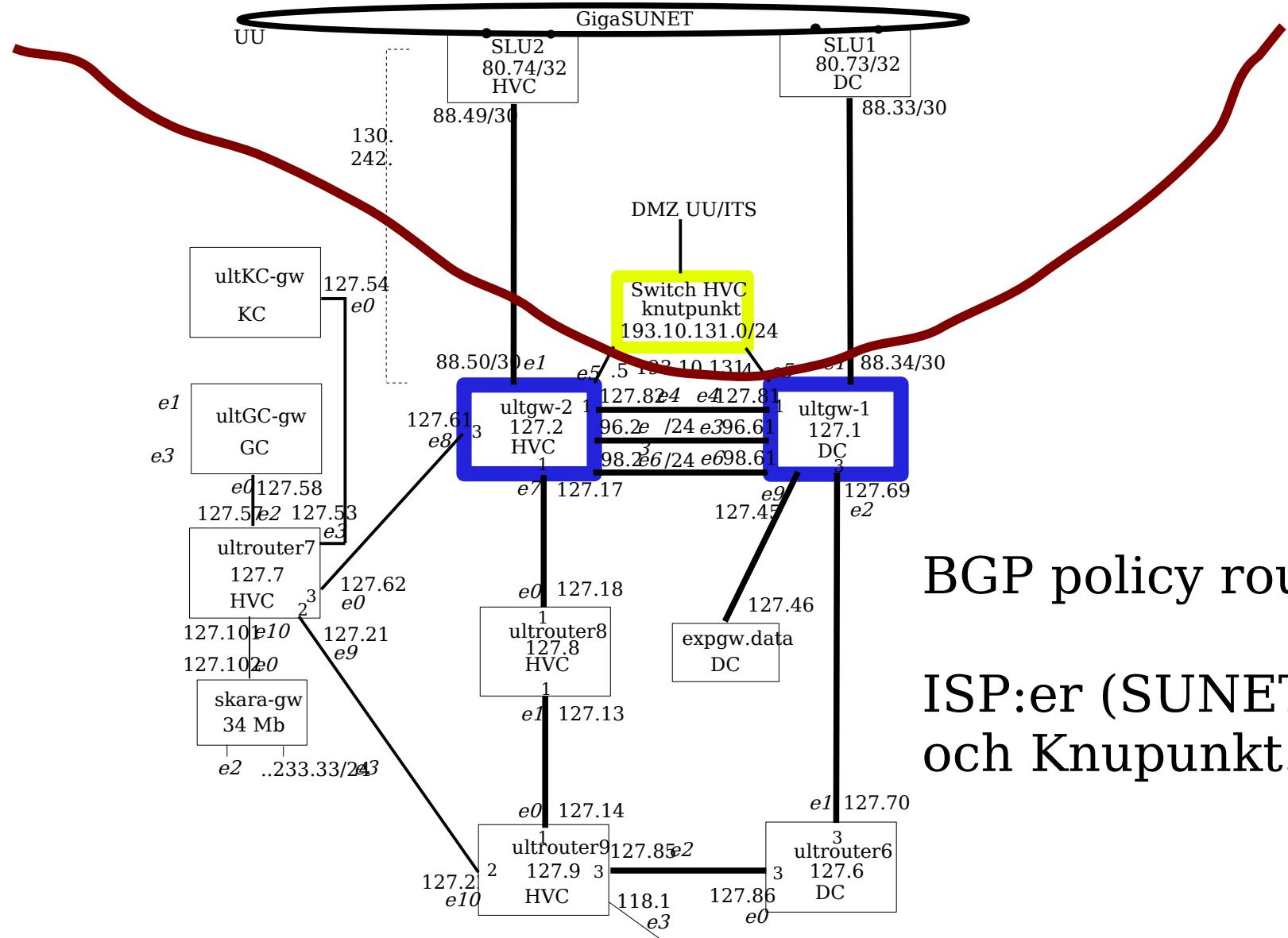
A new network symbol has been seen...

The Penguin Has Landed



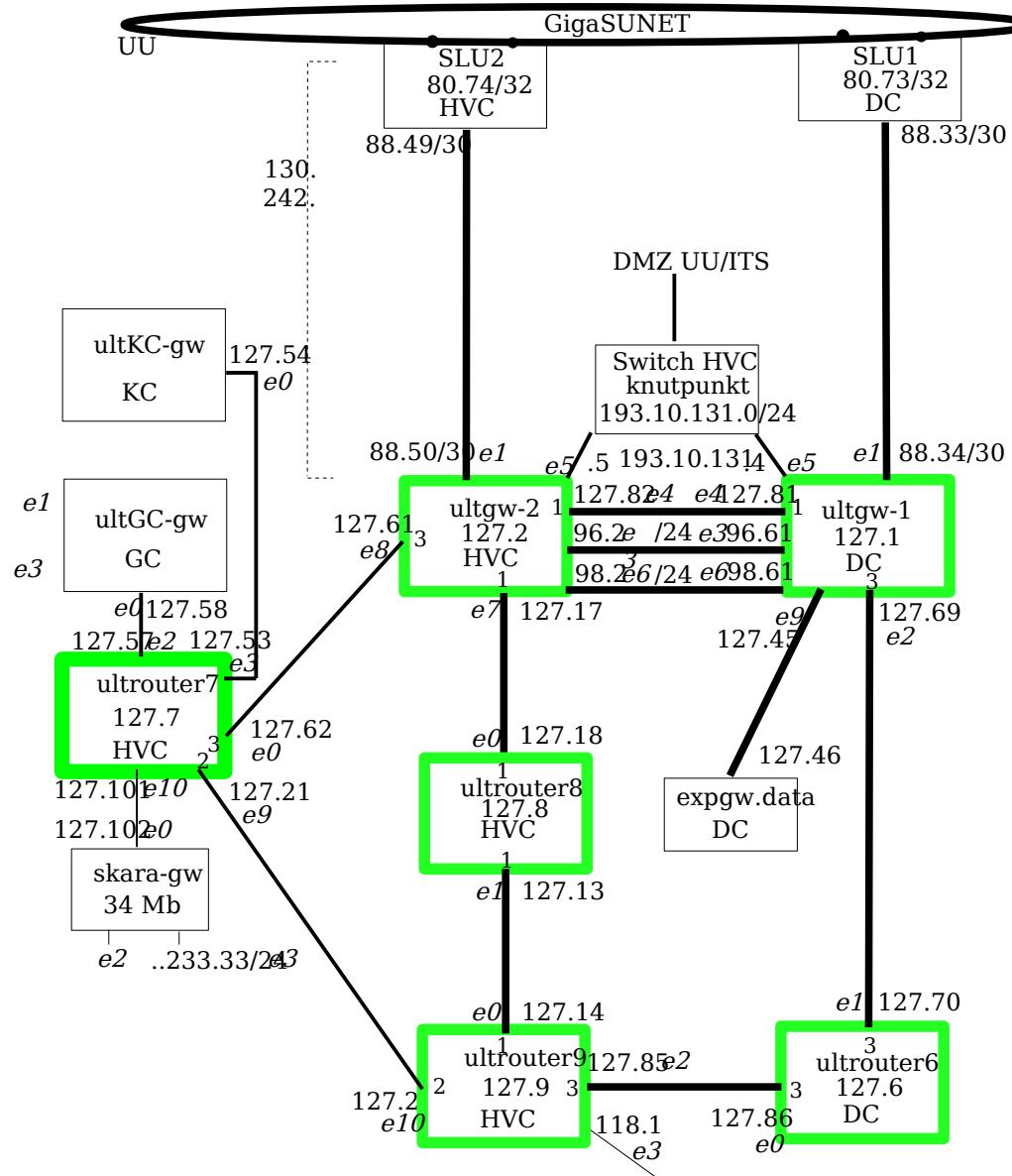
SLU's nät (inte hela)



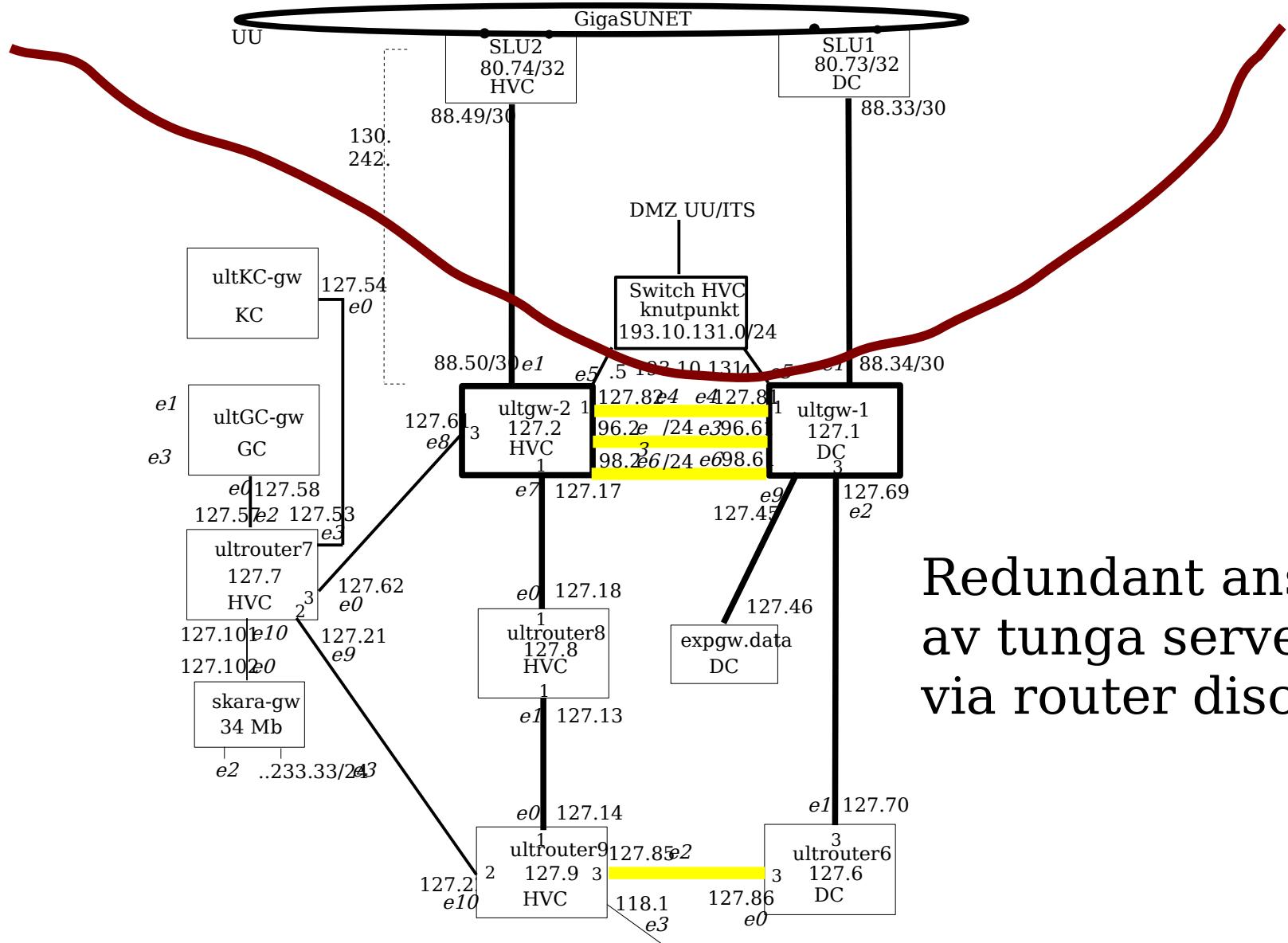


BGP policy routing

ISP:er (SUNET) och Knpunkt.



Redundant
inre kärna



Redundant ansluting
av tunga servernät
via router discovery

