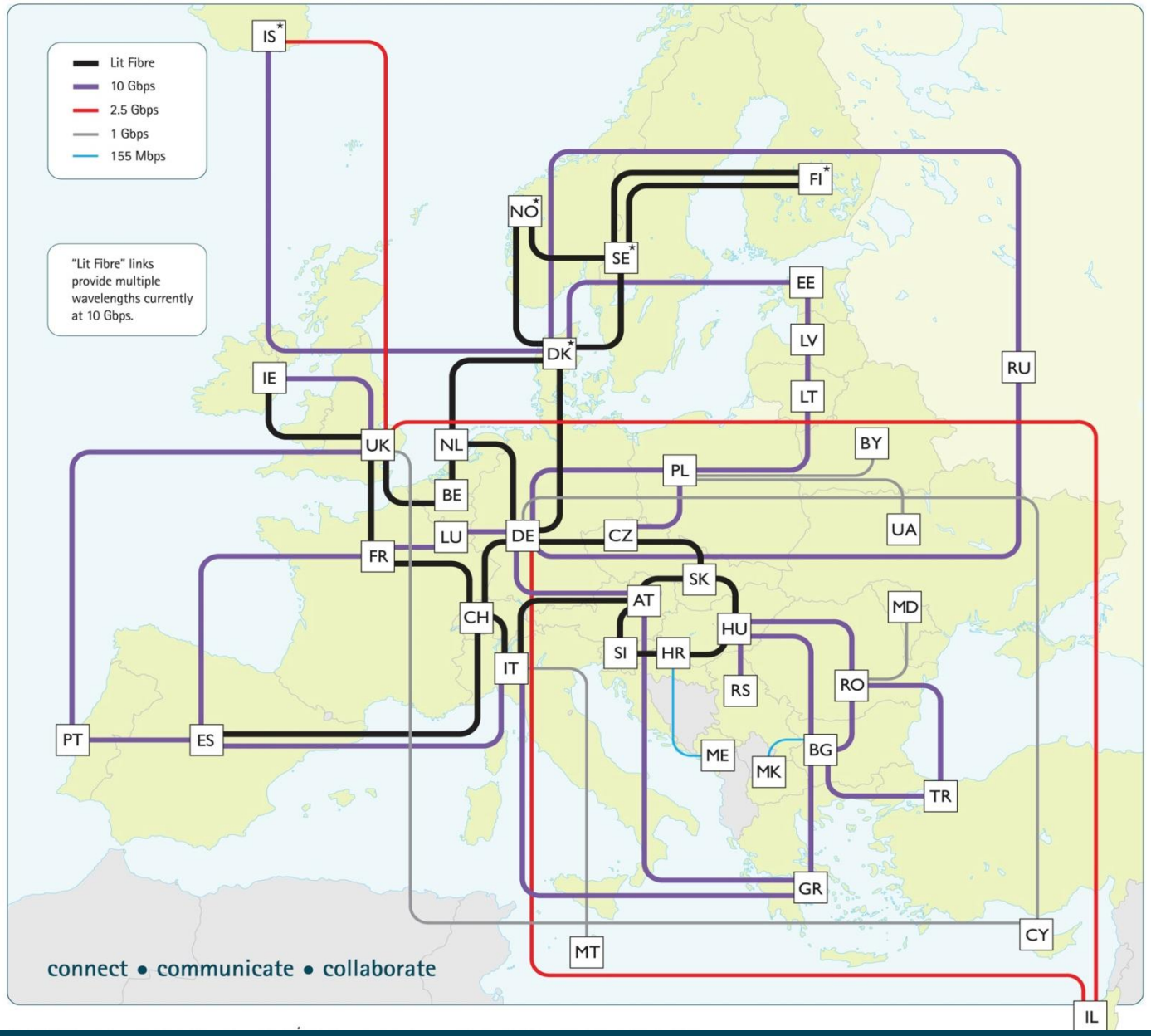
Improving network performance within the cloud: eduPERT and a case study.

Alessandra Scicchitano
SWITCH

Life is full of existential questions:

- Who am I?
- Who do I work for?
- Why am I here?

GEANT


Lit Fibre
10 Gbps
2.5 Gbps
1 Gbps
155 Mbps
"Lit Fibre" links provide multiple wavelengths currently at 10 Gbps.
IS
NO
SE
FI
EE
LV
LT
RU
DK
IE
UK
NL
BE
LU
DE
CZ
PL
BY
UA
FR
SK
AT
CH
IT
SI
HR
HU
MD
RO
RS
PT
ES
ME
MK
BG
TR
GR
CY
MT
IL
connect • communicate • collaborate

I work for

GÉANT

SWITCH

The Swiss Research and Educational network

&

eduPERT

A community of Performance Experts

eduPERT

- eduPERT is an open community to all interested in performance issues
- An extended field of action moving towards the end user and the end point, including positioning for new technologies such as cloud which increase the complexity of performance understanding and diagnosis.
- Increased dissemination of performance expertise among NRENs and end user communities.

How we break boundaries

- Monthly phone calls
- F2F meetings
- PERT KB: http://kb.pert.geant.net/
- eduPERT Portal: http://edupert.gent.net/
- Mailing list: pert-discuss@geant.net (registration link on the portal)
- Performance U!

Performance U!

- “Performance U!” stands for Performance University. Which is an expert school that aims to train performance experts, providing validated performance resources and face to face training.
- “Performance U!” hosts an annual school where the eduPERT community can learn about performance implications of new tools and new technologies and a summer workshop where the community brings outside its experience and knowledge.
- The eduPERT portal (http://edupert.geant.net/) hosts different topics of interest for anyone interested in Performance Enhancement.

eduPERT

Because great performance doesn't know borders or boundaries

SWITCH

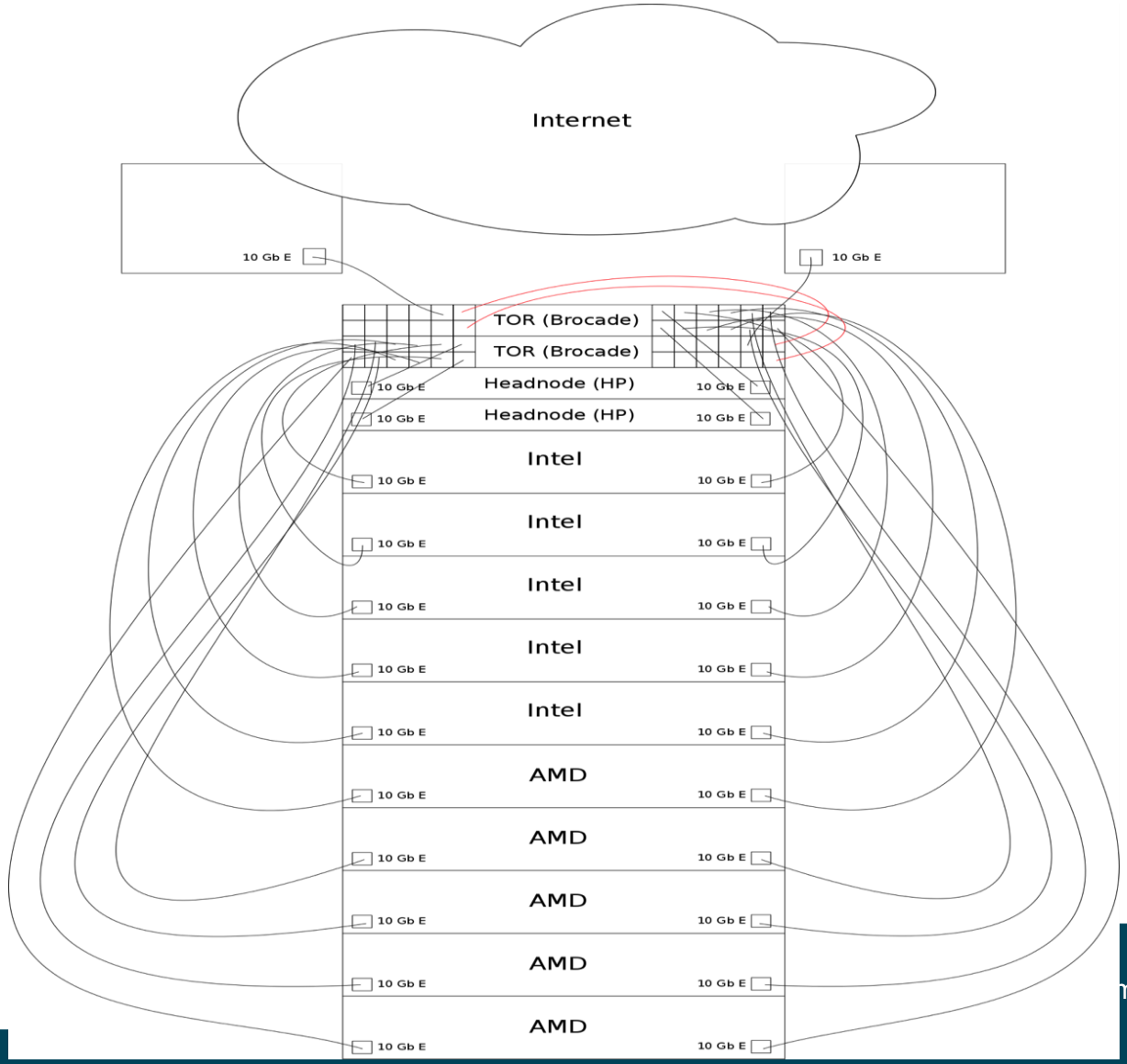
SWITCH

- Happily based where Heidi lives ☺
- A foundation born 26 years ago
- It is the Swiss NREN but it also offers many other services including Cloud

Building Cloud Competence - BCC

- Project on going since September 2012
- The main goal of the project is learning as much as possible about the Cloud and its possible use within our community
- A pilot infrastructure has been built from scratch and it is now open to the community

BCC - Hardware



BCC - Software

- OpenStack
- Puppet
- Ceph-FS
- KVM

The story

- A customer from USI (Universita' della Svizzera Italiana) running his own application within BCC noticed that the max speed he could reach was only 80Mb/s
- Doing some research on his own to understand what the problem could be he read a blog [1] about openstack and high speed network.
- In particular he asked us whether the words "virtio" and "vhost-net" would ring a bell
- It did indeed…

Going back to…

- Almost one year ago: The SWITCH DNS case
- The DNS server is on a VM
- Symptom: Terrible qps (queries per second that the server is able to answer without "dropping")

Before…

Results BIND 9.8.1-P1

	bamus (1 vCPU)			bamus (2 vCPU)			niobe			
	-	+querylog	+querylog, +dsc	-	+querylog	+querylog, +dsc	-	+querylog	+querylog, +dsc	+dsc
manaro	~21'000 qps	~16'000 qps	~10'000 qps	~18'000 qps	~15'000 qps	~10'000 qps	~44'000 qps	~38'000 qps	~28'000 qps	~32'000 qps
asama	~18'500 qps	~15'500 qps	~10'000 qps	~16'000 qps	~14'000 qps	~10'000 qps	~32'000 qps	~31'000 qps	~28'000 qps	~28'000 qps
manaro + asama	~21'000 qps	~16'000 qps	~10'000 qps	~18'000 qps	~15'000 qps	~10'000 qps	***~50'000 qps***	~42'000 qps	~31'000 qps	~35'000 qps

Of course we were immediately able to say where the problem was:

"The bottleneck is located somewhere between the Hardware Interface of the mother host and the BIND (DNS software) in the Virtual machine"

That‘s a joke, right?

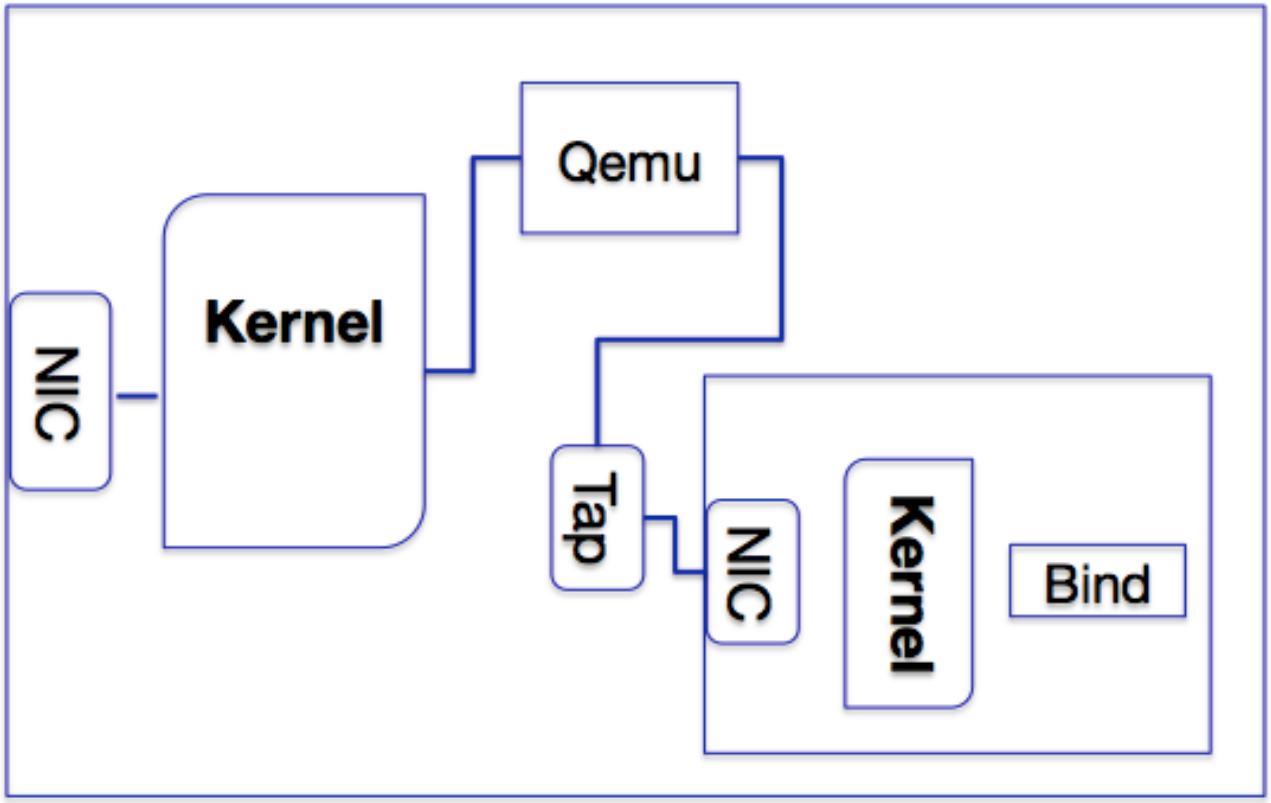
We tried different things before digging into the KVM architecture, but nothing really worked

Virtualization

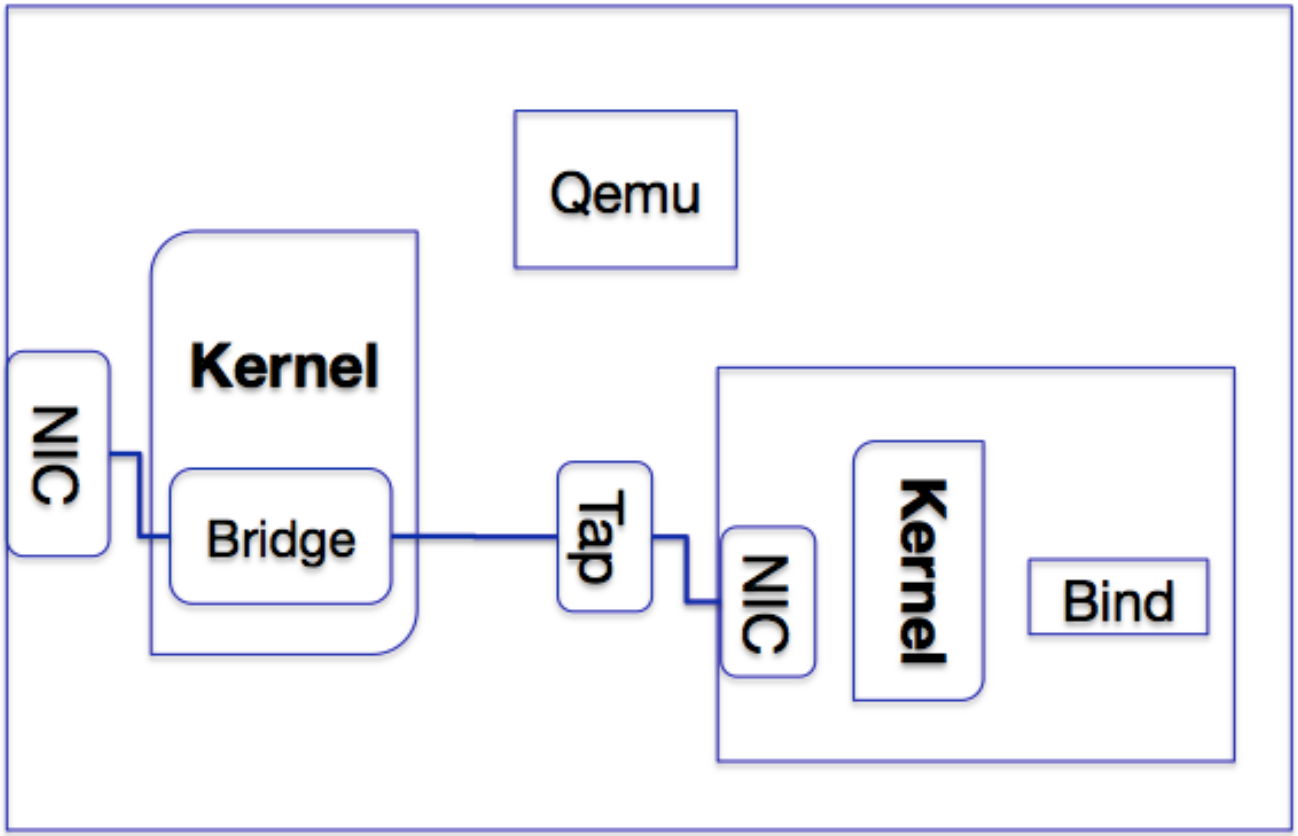
Three different types of virtualization exist:

- Emulation
- Paravirtualization
- Hardware pass-through

KVM - Emulation

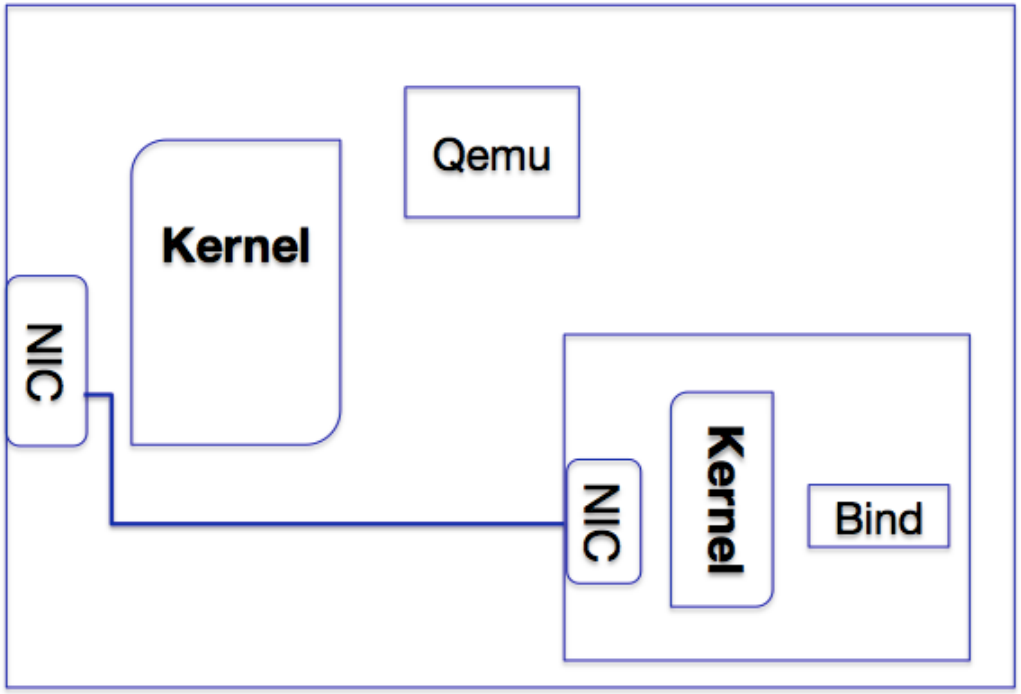


KVM - Paravirtualization



Hardware pass-through

Hardware pass-through means that the guest VM has direct access to the physical hardware.



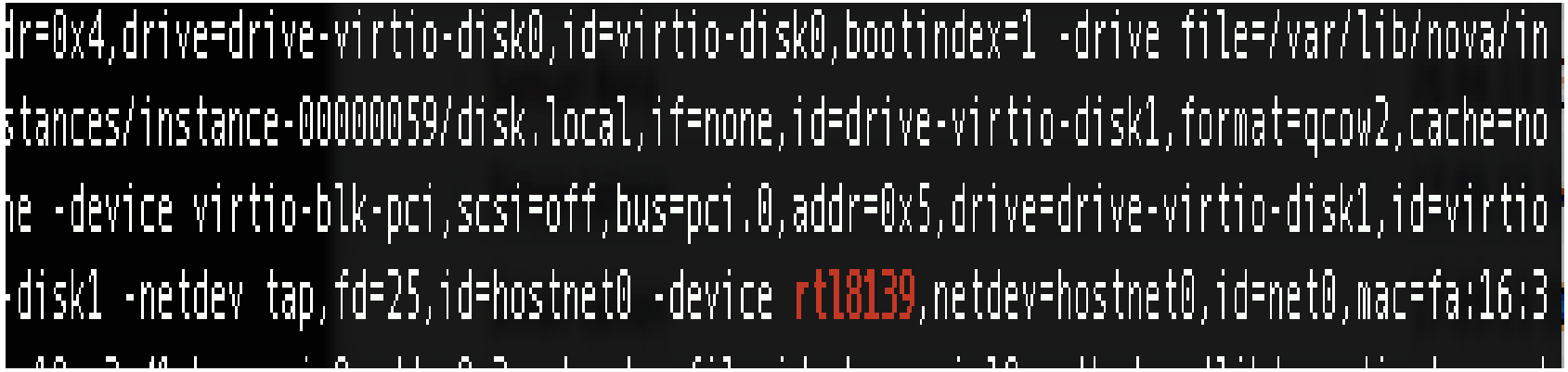
Paravirtualization was the solution to the DNS problem.

After!!!

	simbo (1 vCPU)			simbo (2 vCPU)			simbo (3 vCPU)			
	-	+querylog	+querylog, +dsc	-	+querylog	+querylog, +dsc	-	+querylog	+querylog, +dsc	+dsc
manaro	23'900 qps	16'200 qps	10'700 qps	45'800 qps	28'500 qps	19'500 qps	57'100 qps	36'400 qps	28'300 qps	46'200 qps
manaro + lopevi	27'000 qps	17'700 qps	11'600 qps	**48'800 qps**	29'500 qps	19'500 qps	**73'000 qps**	39'000 qps	30'000 qps	50'300 qps

In BCC

We looked for the type of interface that our VMs had:



Let's paravirtualize BCC

The max speed of the rtl8139 network adapter is 100Mb/s (?????)

No wonder why we couldn't get any better than that.

PARAVIRTUALIZATION!

Howto: Openstack

This line needs to be added in the nova.config file

libvirt_use_virtio_for_bridges=true

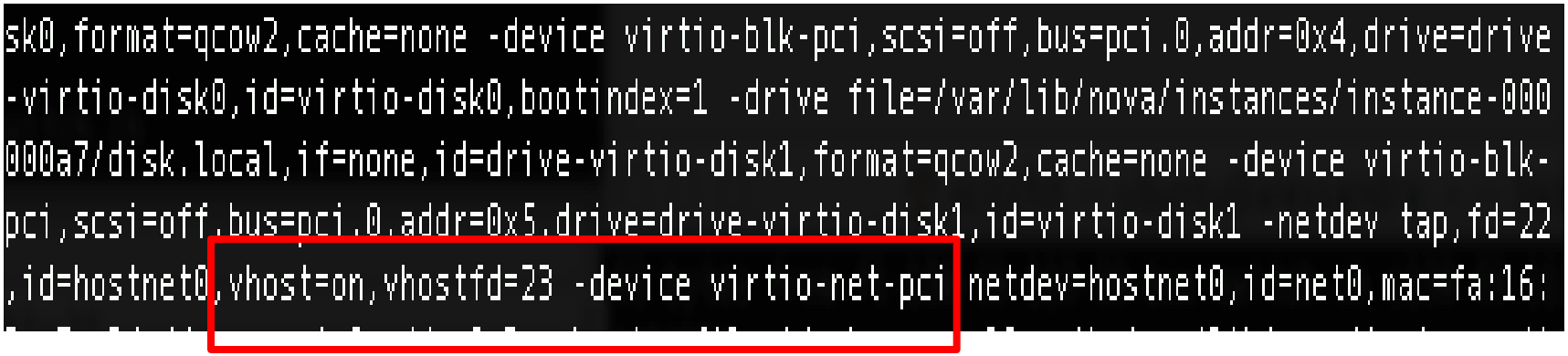
With this modification every VM is now created directly with paravirtualized drivers.

Howto: VMs already existing

Manual steps on the mother-host:

- adding <model type='virtio' /> to the interface in section /var/lib/nova/instances/instance-xxxx/libvirt.xml enables the virtio driver
- virsh destroy instance-xxxxxx
- virsh undefine instance-xxxxxx
- virsh define /var/lib/nova/instance-xxxxxxx/libvirt.xml
- vish start instance-xxxxxx

Drivers now


sk0,format=qcow2,cache=none -device virtio-blk-pci,scsi=off,bus=pci.0,addr=0x4,drive=drive
-virtio-disk0,id=virtio-disk0,bootindex=1 -drive file=/var/lib/nova/instances/instance-000
000a7/disk.local,if=none,id=drive-virtio-disk1,format=qcow2,cache=none -device virtio-blk-
pci,scsi=off,bus=pci.0,addr=0x5,drive=drive-virtio-disk1,id=virtio-disk1 -netdev tap,fd=22
,id=hostnet0,vhost=on,vhostfd=23 -device virtio-net-pci netdev=hostnet0,id=net0,mac=fa:16:

Comparison

- RTL8139

```
root@questnet:~# iperf -i 1 -c 10.0.0.20
------------------------------------------------------------
Client connecting to 10.0.0.20, TCP port 5001
TCP window size: 22.9 KByte (default)
------------------------------------------------------------
[  3] local 10.0.0.29 port 57754 connected with 10.0.0.20 port 5001
[ ID] Interval       Transfer     Bandwidth
[  3]  0.0- 1.0 sec  14.2 MBytes   120 Mbits/sec
[  3]  1.0- 2.0 sec  16.4 MBytes   137 Mbits/sec
[  3]  2.0- 3.0 sec  16.9 MBytes   142 Mbits/sec
[  3]  3.0- 4.0 sec  15.4 MBytes   129 Mbits/sec
[  3]  4.0- 5.0 sec  14.1 MBytes   118 Mbits/sec
[  3]  5.0- 6.0 sec  15.9 MBytes   133 Mbits/sec
[  3]  6.0- 7.0 sec  15.5 MBytes   130 Mbits/sec
[  3]  7.0- 8.0 sec  16.4 MBytes   137 Mbits/sec
[  3]  8.0- 9.0 sec  16.1 MBytes   135 Mbits/sec
[  3]  9.0-10.0 sec  16.9 MBytes   142 Mbits/sec
[  3]  0.0-10.0 sec   158 MBytes   132 Mbits/sec
root@questnet:~# owping 10.0.0.4
```

Comparison

- First improvement using e1000

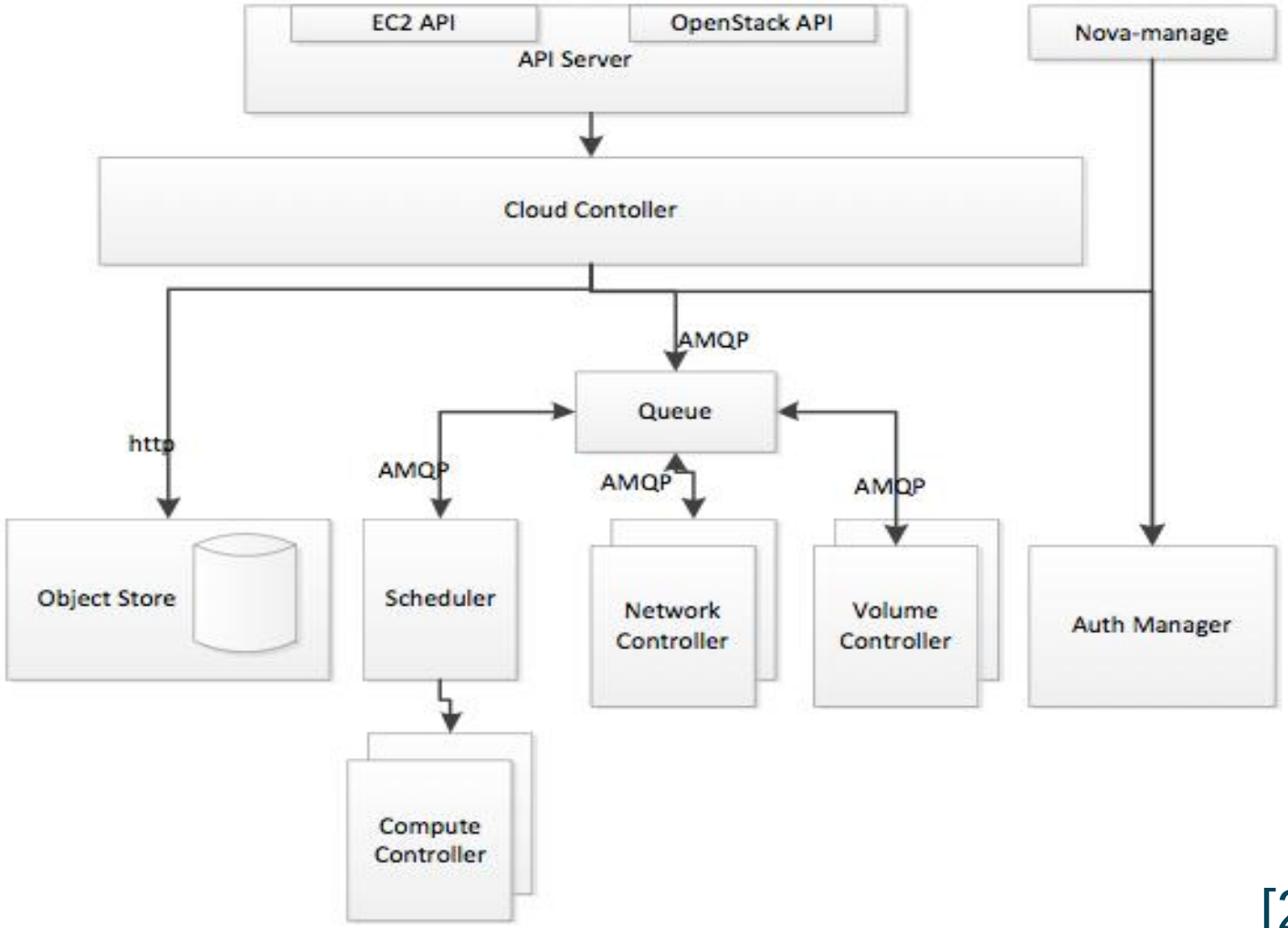
```
ubuntu@questnet:~$ iperf -c 10.0.0.4 -i 1
------------------------------------------------------------
Client connecting to 10.0.0.4, TCP port 5001
TCP window size: 22.9 KByte (default)
------------------------------------------------------------
[  3] local 10.0.0.29 port 48665 connected with 10.0.0.4 port 5001
[ ID] Interval       Transfer     Bandwidth
[  3]  0.0- 1.0 sec  11.6 MBytes  97.5 Mbits/sec
[  3]  1.0- 2.0 sec  16.1 MBytes   135 Mbits/sec
[  3]  2.0- 3.0 sec  28.9 MBytes   242 Mbits/sec
[  3]  3.0- 4.0 sec  46.0 MBytes   386 Mbits/sec
[  3]  4.0- 5.0 sec  40.5 MBytes   340 Mbits/sec
[  3]  5.0- 6.0 sec  25.5 MBytes   214 Mbits/sec
[  3]  6.0- 7.0 sec  29.9 MBytes   251 Mbits/sec
[  3]  7.0- 8.0 sec  30.4 MBytes   255 Mbits/sec
[  3]  8.0- 9.0 sec  29.6 MBytes   249 Mbits/sec
[  3]  9.0-10.0 sec  37.5 MBytes   315 Mbits/sec
[  3]  0.0-10.0 sec   296 MBytes   248 Mbits/sec
ubuntu@questnet:~$
```

Comparison

- VIRTIO

```
ubuntu@questnet:~$ iperf -c 10.0.0.4 -i 1
------------------------------------------------------------
Client connecting to 10.0.0.4, TCP port 5001
TCP window size: 22.9 KByte (default)
------------------------------------------------------------
[  3] local 10.0.0.29 port 48661 connected with 10.0.0.4 port 5001
[ ID] Interval       Transfer     Bandwidth
[  3]  0.0- 1.0 sec   849 MBytes  7.12 Gbits/sec
[  3]  1.0- 2.0 sec   937 MBytes  7.86 Gbits/sec
[  3]  2.0- 3.0 sec   876 MBytes  7.34 Gbits/sec
[  3]  3.0- 4.0 sec   857 MBytes  7.19 Gbits/sec
[  3]  4.0- 5.0 sec   935 MBytes  7.84 Gbits/sec
[  3]  5.0- 6.0 sec   956 MBytes  8.02 Gbits/sec
[  3]  6.0- 7.0 sec   972 MBytes  8.15 Gbits/sec
[  3]  7.0- 8.0 sec   976 MBytes  8.18 Gbits/sec
[  3]  8.0- 9.0 sec   956 MBytes  8.02 Gbits/sec
[  3]  9.0-10.0 sec   991 MBytes  8.32 Gbits/sec
[  3]  0.0-10.0 sec  9.09 GBytes  7.81 Gbits/sec
ubuntu@questnet:~$
```

OpenStack Components



[2]

References

[1] http://buriedlede.blogspot.ch/2012/11/driving-100-gigabit-network-with.html

[2] http://salsahpc.indiana.edu/b534projects/sites/default/files/public/6_Behind%20the%20scenes%20of%20IaaS%20implementations_AlRwais_Sumayah%20Abdulaziz.pdf