Infiniband

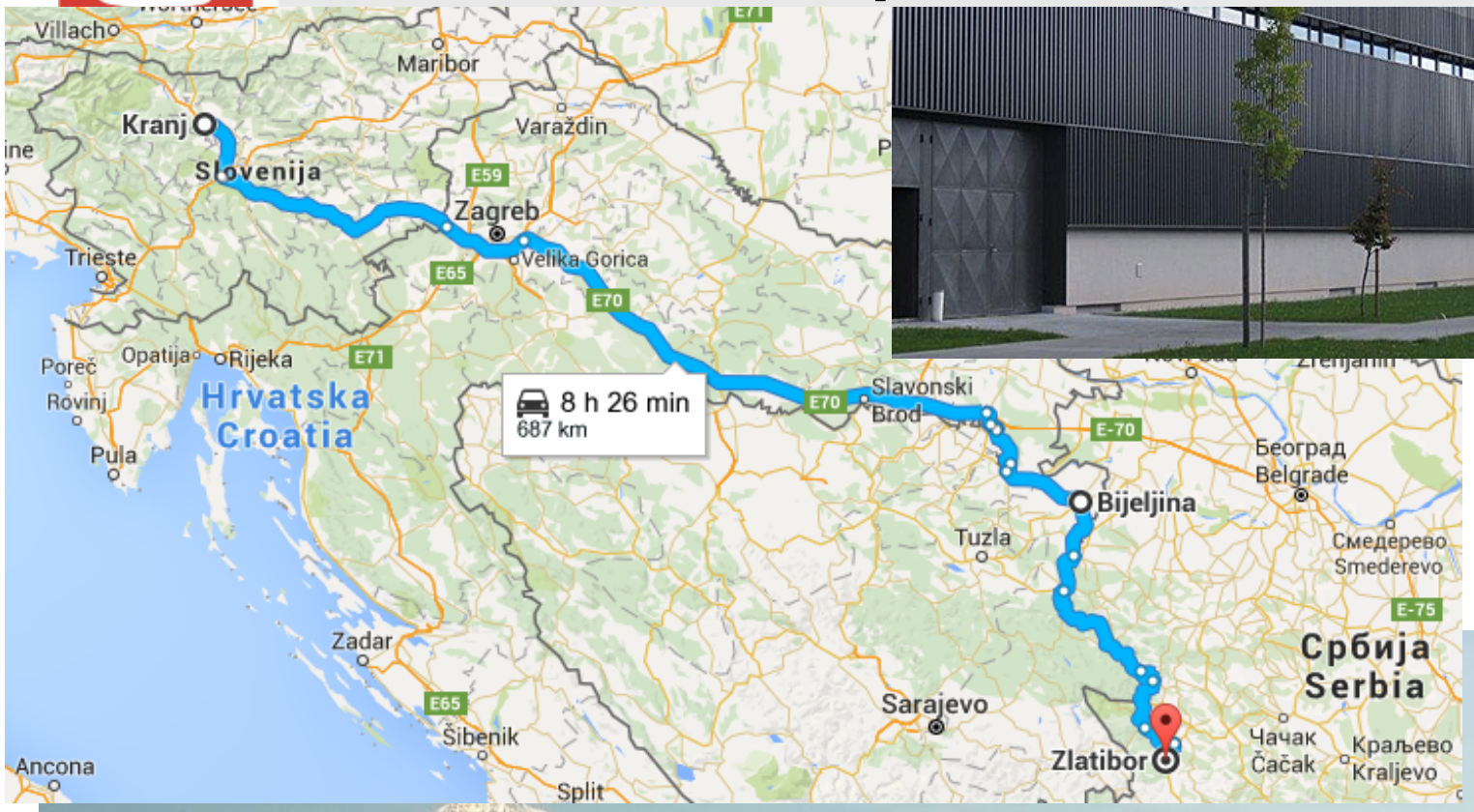
- **mag. Sergej Rožman**; Abakus plus d.o.o.
- The latest version of this document is available at: http://www.abakus.si/

Infiniband

mag. Sergej Rožman

sergej.rozman@abakus.si

Abakus plus d.o.o. - Kranj



Abakus plus d.o.o.

ORACLE Gold Partner

History

- from 1992, ~20 employees

Applications:

- special (DB – Newspaper Distribution, FIS – Flight Information System)
- **ARBITER – the ultimate tool in audit trailing**
- **APPM - Abakus Plus Performance Monitoring Tool**

Services:

- DBA, OS administration , programming (MediaWiki, Oracle)
- networks (services, VPN, QoS, security)
- open source, monitoring (Nagios, OCS, Wiki)

Infrastructure:

- servers, **SAN storage**, firewalls, **backup servers**

Infrastructure:

- from 1995 GNU/Linux ***(~20 years of experience !)***
- Oracle on GNU/Linux: since RDBMS 7.1.5 & Forms 3.0 ***(before Oracle !)***
- **>25 years of experience with High-Availability !**

Mestna občina Ljubljana

MESTNA OBČINA KOPER
COMUNE CITTA DI CAPODISTRIA

Iskra

Aerodrom Ljubljana

KONTROLA ZRAČNEGA PROMETA SLOVENIJE

NOVA BANKA

GOODYEAR

HARDWARE

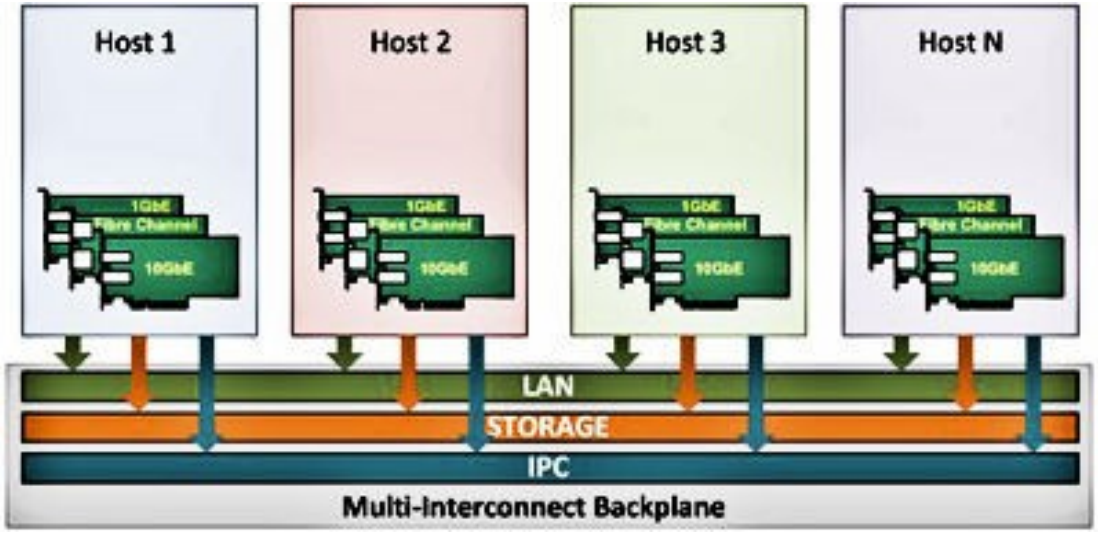
AS

Abakus

As na disku.

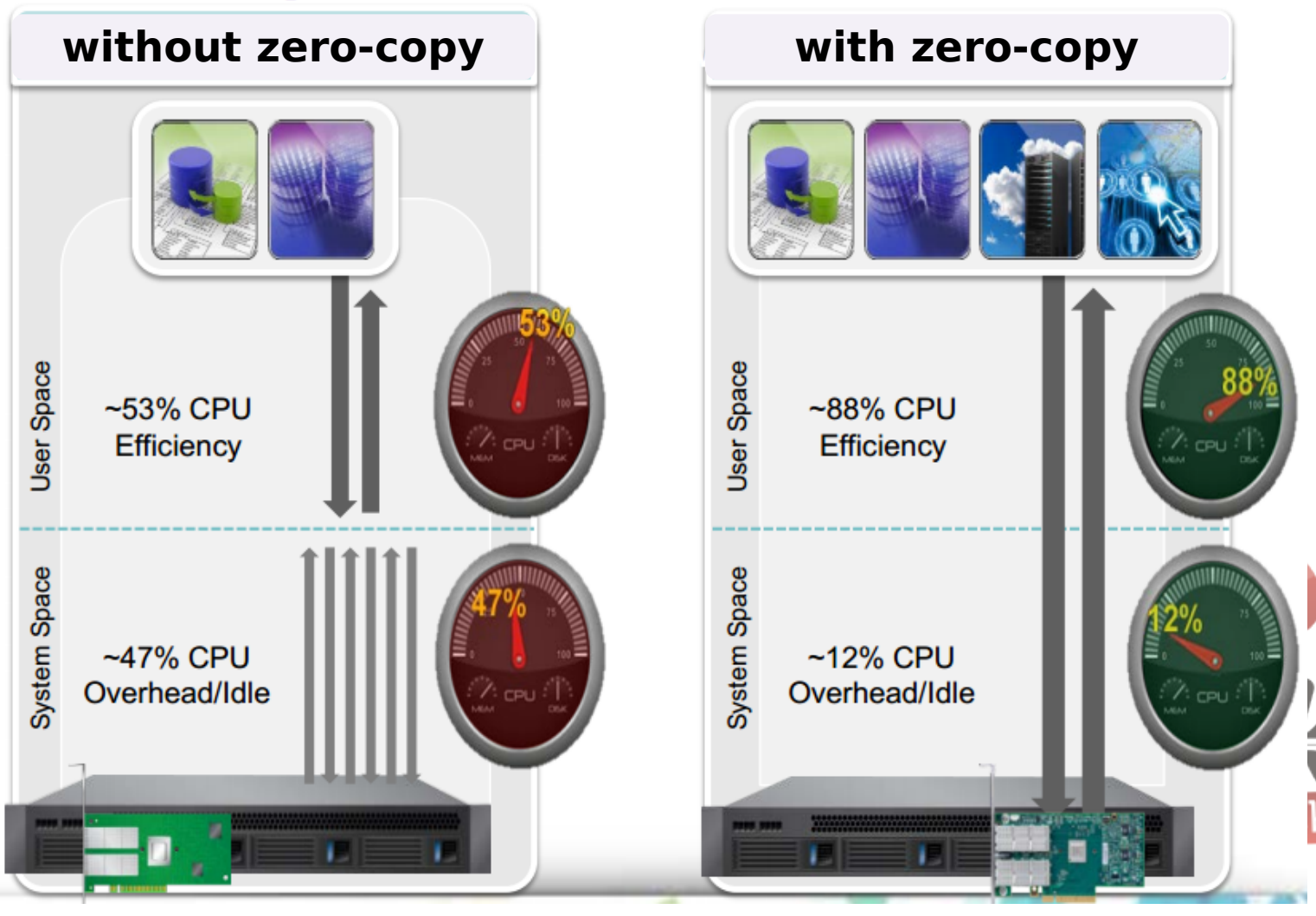
What is Infiniband

- high throughput
- low end-to-end latency < 2 µs (1 GbE: ~35 µs)
- RDMA
- The idea was to provide a single switched fabric that would link computers and storage

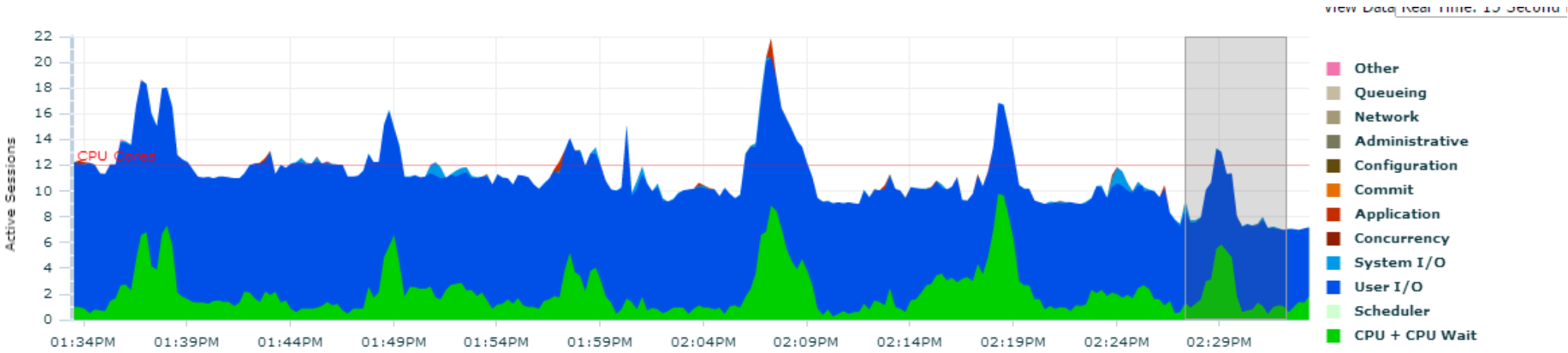


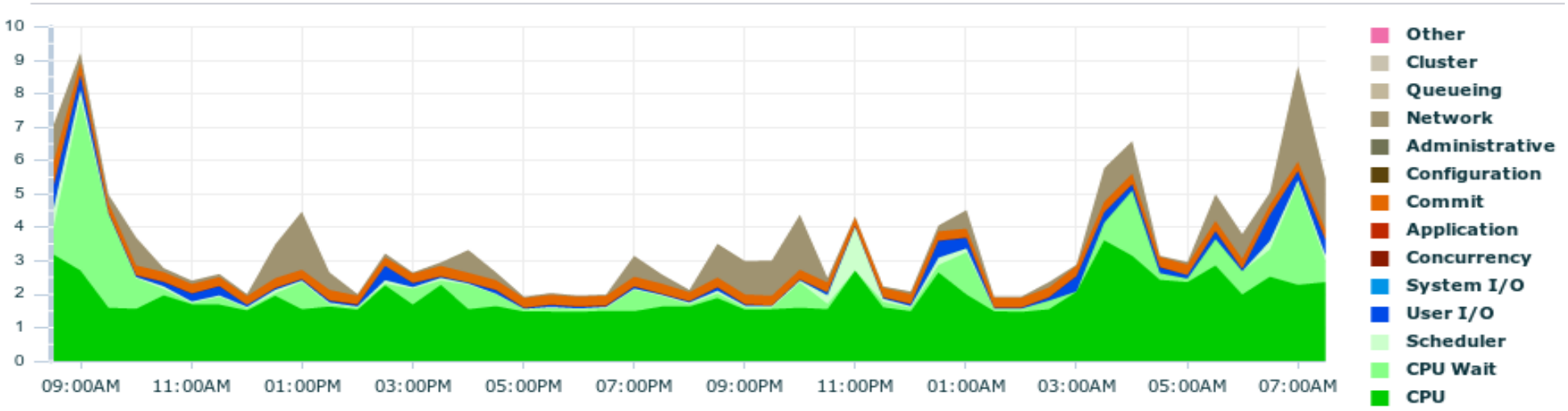
IB RDMA

- Remote Direct Memory Access
- Zero-copy

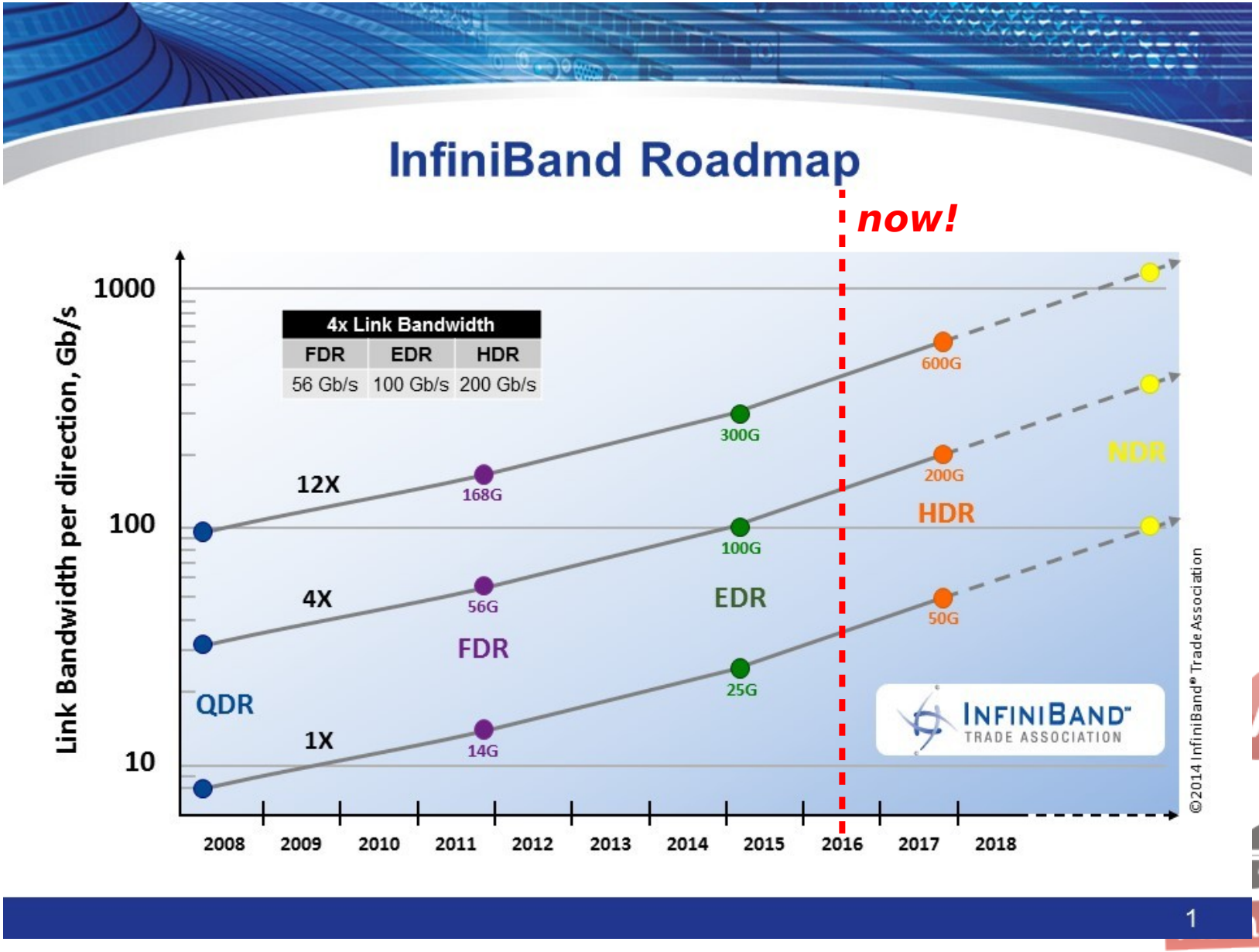


IB Transformation





IB Roadmap



source: Infiniband Trade Association; http://www.infinibandta.org/content/pages.php?pg=technology_overview

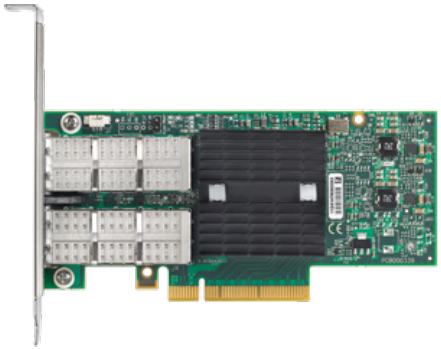
Infiniband Main Players

- **Mellanox**, (Yokneam, Israel)
  Since 2010, Oracle Corporation has been a major investor in the company, holding around 10% of its stock. Oracle uses its InfiniBand technology in its **Exadata** and **Exalogic** appliances.

- **Intel**
  In 2012, Intel acquires InfiniBand business from Qlogic
  Rumors: Infiniband will be eventually integrated directly into CPU.

AS
Abakus
As na disku.

Hardware



- HCA – Host Channel Adapter

- Switch
  Mellanox: 8 ports to 648 ports

- Fibre and Copper cables
  with QSFP connectors

- Long-Haul System
  distance: 1km to 80km
  latency: 200ns + 5µs/km

SOFTWARE

AS

Abakus

As na disku.

Software

- De facto standard software developed by OpenFabrics Alliance. OFED (OpenFabrics Enterprise Distribution)
  - Linux (and FreeBSD)
  - Microsoft Windows

Oracle Linux 7 (x86_64) OFED 2.0

Latest OFED 2.0 packages for Oracle Linux 7 (x86_64)

```
/etc/yum.repos.d/public-yum-ol7.repo

[ol7_UEKR3_OFED20]
name=OFED supporting tool packages for Unbreakable Enterprise Kernel on Oracle Linux 7 ($basearch)
baseurl=http://public-yum.oracle.com/repo/OracleLinux/OL7/UEKR3_OFED20/$basearch/
gpgkey=file:///etc/pki/rpm-gpg/RPM-GPG-KEY-oracle
gpgcheck=1
enabled=1
priority=20
```

Basic IB* Commands

 # ibhosts

```
Ca      : 0x0002c903002b8cb2 ports 1 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0x0002c90300108978 ports 1 "lonhbs2 mlx4_0"
Ca      : 0x0002c903005ad624 ports 1 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0xf452140300e1a676 ports 1 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0x0002c903005ad638 ports 1 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0x0002c903002c561e ports 1 "ebtestnovi mlx4_0"
Ca      : 0x0002c9030056318e ports 1 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0x0002c903004cb2bc ports 1 "lonhbs1 mlx4_0"
```

 # ibswitches

```
Switch  : 0x0002c9020046fc68 ports 8 "Infiniscale-IV Mellanox Technologies" base port 0 lid 2 lmc 0
Switch  : 0x0002c902004707b8 ports 8 "Infiniscale-IV Mellanox Technologies" base port 0 lid 4 lmc 0
```

Basic IB* Commands

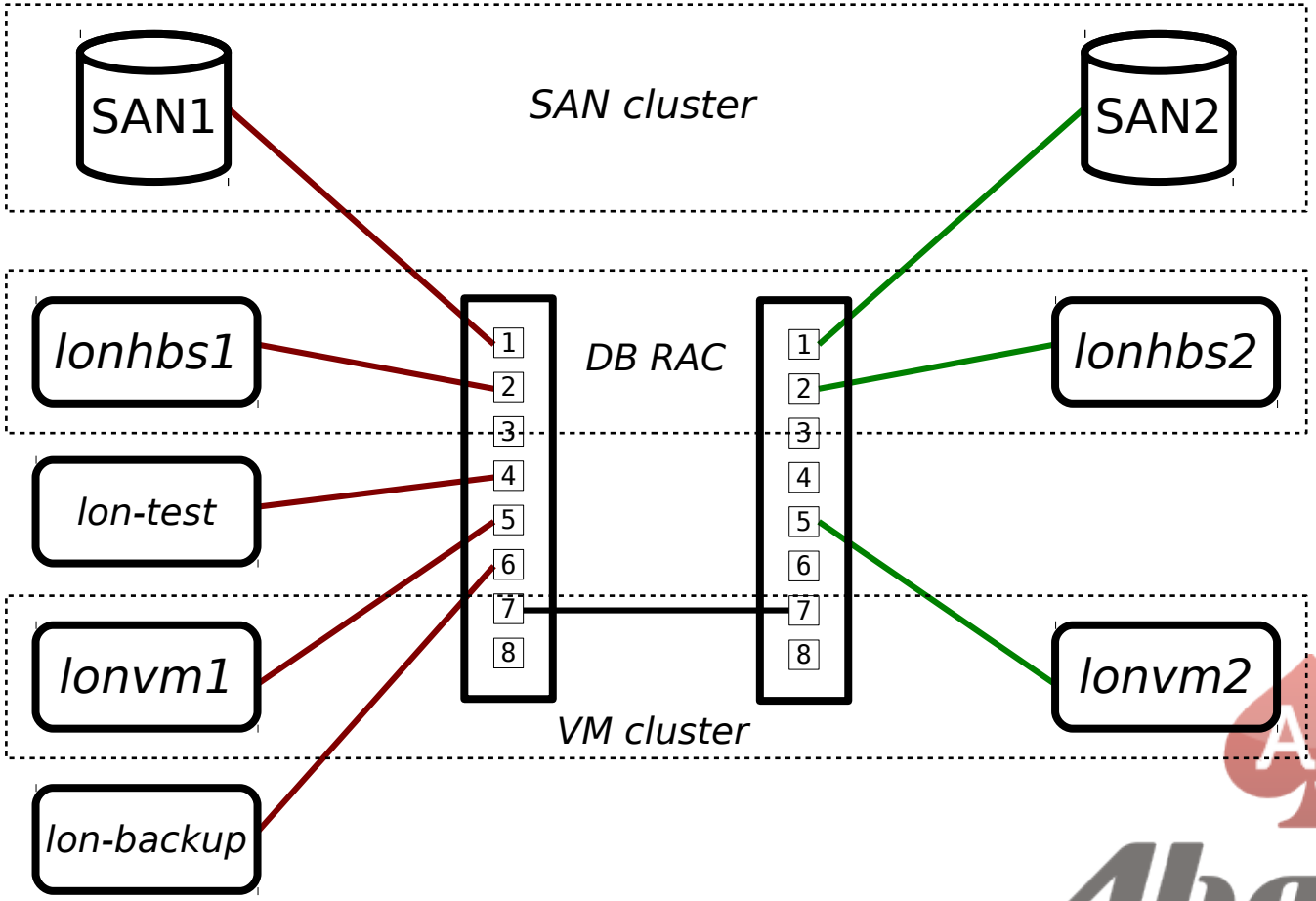
```
# ibnodes
Ca      : 0x0002c903002b8cb2 ports 1 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0x0002c903005ad624 ports 1 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0x0002c90300108978 ports 1 "lonhbs2 mlx4_0"
Ca      : 0x0002c903005ad638 ports 1 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0xf452140300e1a676 ports 1 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0x0002c903002c561e ports 1 "ebtestnovi mlx4_0"
Ca      : 0x0002c9030056318e ports 1 "MT25408 ConnectX Mellanox Technologies"
Ca      : 0x0002c903004cb2bc ports 1 "lonhbs1 mlx4_0"
Switch  : 0x0002c9020046fc68 ports 8 "Infiniscale-IV Mellanox Technologies" base port 0 lid 2 lmc 0
Switch  : 0x0002c902004707b8 ports 8 "Infiniscale-IV Mellanox Technologies" base port 0 lid 4 lmc 0
```

Basic IB* Commands

iblinkinfo

```
CA: MT25408 ConnectX Mellanox Technologies:
      0x0002c903002b8cb3      9    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       2    5[  ] "Infiniscale-IV Mellanox Technologies" ( )
CA: MT25408 ConnectX Mellanox Technologies:
      0x0002c903005ad625      1    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       2    1[  ] "Infiniscale-IV Mellanox Technologies" ( )
CA: lonhbs2 mlx4_0:
      0x0002c90300108979      5    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       2    2[  ] "Infiniscale-IV Mellanox Technologies" ( )
Switch: 0x0002c9020046fc68 Infiniscale-IV Mellanox Technologies:
           2    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       1    1[  ] "MT25408 ConnectX Mellanox Technologies" ( )
           2    2[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       5    1[  ] "lonhbs2 mlx4_0" ( )
           2    3[  ] ==(                Down/ Polling)==>             [  ] "" ( )
           2    4[  ] ==(                Down/ Polling)==>             [  ] "" ( )
           2    5[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       9    1[  ] "MT25408 ConnectX Mellanox Technologies" ( )
           2    6[  ] ==(                Down/ Polling)==>             [  ] "" ( )
           2    7[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       4    7[  ] "Infiniscale-IV Mellanox Technologies" ( )
           2    8[  ] ==(                Down/ Polling)==>             [  ] "" ( )
CA: MT25408 ConnectX Mellanox Technologies:
      0xf452140300e1a677     11    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       4    6[  ] "Infiniscale-IV Mellanox Technologies" ( )
CA: MT25408 ConnectX Mellanox Technologies:
      0x0002c903005ad639      7    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       4    5[  ] "Infiniscale-IV Mellanox Technologies" ( )
CA: lonhbs1 mlx4_0:
      0x0002c903004cb2bd      6    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       4    2[  ] "Infiniscale-IV Mellanox Technologies" ( )
CA: MT25408 ConnectX Mellanox Technologies:
      0x0002c9030056318f      3    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       4    1[  ] "Infiniscale-IV Mellanox Technologies" ( )
Switch: 0x0002c902004707b8 Infiniscale-IV Mellanox Technologies:
           4    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       3    1[  ] "MT25408 ConnectX Mellanox Technologies" ( )
           4    2[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       6    1[  ] "lonhbs1 mlx4_0" ( )
           4    3[  ] ==(                Down/ Polling)==>             [  ] "" ( )
           4    4[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>      10    1[  ] "ebtestnovi mlx4_0" ( )
           4    5[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       7    1[  ] "MT25408 ConnectX Mellanox Technologies" ( )
           4    6[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>      11    1[  ] "MT25408 ConnectX Mellanox Technologies" ( )
           4    7[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       2    7[  ] "Infiniscale-IV Mellanox Technologies" ( )
           4    8[  ] ==(                Down/ Polling)==>             [  ] "" ( )
CA: ebtestnovi mlx4_0:
      0x0002c903002c561f     10    1[  ] ==( 4X          10.0 Gbps Active/  LinkUp)==>       4    4[  ] "Infiniscale-IV Mellanox Technologies" ( )
```

IB Topology Example



Diagnostic IB* Commands

```
# ibqueryerrors
Errors for "MT25408 ConnectX Mellanox Technologies"
   GUID 0x2c903002b8cb3 port 1: [PortXmitWait == 10455028]
Errors for "MT25408 ConnectX Mellanox Technologies"
   GUID 0x2c903005ad625 port 1: [PortXmitWait == 4294967295]
Errors for "lonhbs2 mlx4_0"
   GUID 0x2c90300108979 port 1: [PortXmitWait == 20342674]
Errors for "MT25408 ConnectX Mellanox Technologies"
   GUID 0xf452140300e1a677 port 1: [PortXmitWait == 5277]
Errors for 0x2c9020046fc68 "Infiniscale-IV Mellanox Technologies"
   GUID 0x2c9020046fc68 port ALL: [PortXmitWait == 902127241]
   GUID 0x2c9020046fc68 port 1: [PortXmitWait == 4294967295]
   GUID 0x2c9020046fc68 port 2: [PortXmitWait == 4294967295]
   GUID 0x2c9020046fc68 port 3: [PortXmitWait == 3896593]
   GUID 0x2c9020046fc68 port 5: [PortXmitWait == 898230651]
   GUID 0x2c9020046fc68 port 7: [PortXmitWait == 4294967295]
Errors for "MT25408 ConnectX Mellanox Technologies"
   GUID 0x2c903005ad639 port 1: [PortXmitWait == 13079]
Errors for "lonhbs1 mlx4_0"
   GUID 0x2c903004cb2bd port 1: [PortXmitWait == 32758967]
Errors for "MT25408 ConnectX Mellanox Technologies"
   GUID 0x2c9030056318f port 1: [PortXmitWait == 4294967295]
...
```

AS Abakus As na disku.

Diagnostic IB* Commands

 # perfquery

```
# Port counters: Lid 10 port 1 (CapMask: 0x1400)
PortSelect:......................1
CounterSelect:...................0x0000
SymbolErrorCounter:..............0
LinkErrorRecoveryCounter:........0
LinkDownedCounter:...............0
PortRcvErrors:...................0
PortRcvRemotePhysicalErrors:.....0
PortRcvSwitchRelayErrors:........0
PortXmitDiscards:................0
PortXmitConstraintErrors:........0
PortRcvConstraintErrors:.........0
CounterSelect2:..................0x00
LocalLinkIntegrityErrors:........0
ExcessiveBufferOverrunErrors:....0
VL15Dropped:.....................0
PortXmitData:....................4294967295
PortRcvData:.....................4294967295
PortXmitPkts:....................592243248
PortRcvPkts:.....................699785630
PortXmitWait:....................4294967295
```

USAGE

AS

Abakus

As na disku.

IP over IB

- TCP/IP over infiniband network
  (kernel module: ib_ipoib)

```
# ifconfig ib0
ib0       Link encap:UNSPEC  HWaddr 80-00-00-48-FE-80-00-00-00-00-00-00-00-00-00-00
          inet addr:172.16.25.112  Bcast:172.16.25.255  Mask:255.255.255.0
          inet6 addr: fe80::202:c903:5a:d625/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:65520  Metric:1
          RX packets:6170810002 errors:0 dropped:6 overruns:0 frame:0
          TX packets:6825829184 errors:0 dropped:101 overruns:0 carrier:0
          collisions:0 txqueuelen:256
          RX bytes:23241091268923 (23.2 TB)  TX bytes:127124204361366 (127.1 TB)
```

- Unreliable Datagram (UD) mode or Connected mode (CM)

```
echo datagram > /sys/class/net/ib0/mode
echo connected > /sys/class/net/ib0/mode
```

- RFC4392 - IP over InfiniBand (IPoIB) Architecture
  RFC4391 - Transmission of IP over InfiniBand (IPoIB) (UD mode)
  RFC4755 - IP over InfiniBand: Connected Mode

Cluster Interconnect (RAC)

- traditional IP/UDP (still possible – IP over IB)
  (kernel module: ib_ipoib)

- native IB RDS (Reliable Datagram Sockets)
  (kernel module: rds_rdma.ko, rds.ko)

- My Oracle Support
  How to Install CRS and RAC with Infiniband HCAs and RDS Protocol **(Doc ID 943025.1)**

```
relink the binaries for RDS on all nodes in the RDBMS homes (optional ASM homes):
1. change into ORACLE_HOME/rdbms/lib
2. set the ORACLE_HOME environment variable
3. link with: make -f ins_rdbms.mk ipc_rds ioracle
```

SAN

- SRP (SCSI RDMA protocol)
- iSER (iSCSI Extensions for RDMA)
- iSCSI (IP over IB)

- initiator (client): built into the kernel
- target (server): SCST, LIO, STGT, IET

AS
Abakus
As na disku.

NAS (NFS over RDMA)

- Server

```
modprobe svcrdma
echo rdma 20049 > /proc/fs/nfsd/portlist
```

```
/etc/exports:
/directory        172.16.25.107(fsid=1,rw,sync,insecure,no_subtree_check)
```

- Client

```
mount -o rdma,port=20049 172.16.25.114:/directory /mnt
```

Abakus
As na disku.

Other Services

- Distributed Replicated Block Device (DRBD)
  (network based mirror/RAID)
- distributed file systems (GlusterFS, Lustre)
- ibping, ibtracert
- simple services (file transfer etc.) ?

Performance

- test 1 (notebook with SSD, DB on VM):
  max_iops = **9.983**, latency = **8**, max_mbps = **251**
- test 2 (test DB, 10x 600 GB 15k FC):
  max_iops = **1.824**, latency = **11**, max_mbps = **280**
- test 3 (production DB, 30x 146 GB 15k FC):
  max_iops = **6.498**, latency = **10**, max_mbps = **455**
- test 4 (**Abakus SAN**, 16x SSD, **Infiniband 40G**):
  max_iops = **43.782**, latency = **0**, max_mbps = **1.727**

Abakus
As na disku.

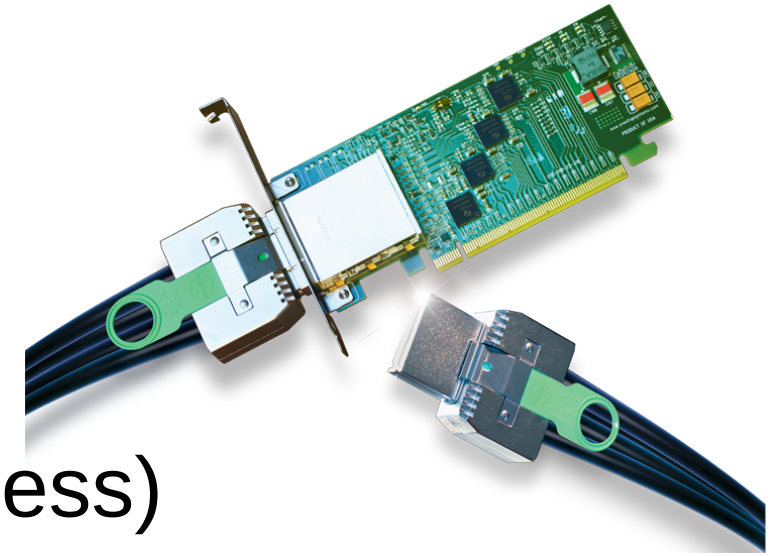
Price

~ in EUR	HBA	switch 8-port	cable
10 GbE	500	1.000	5 (cat. 6a)
FC (8 Gbit/s)	500	2.500	f:50 + t:100
Infiniband (40 Gbit/s)	500	2.000	c:100 f:200

AS Abakus As na disku.

Future Directions

- Infiniband
- Ethernet
- PCIe over Cable (PCI Express)
- ... ?



Infiniband

Thank You

mag. Sergej Rožman
ABAKUS plus d.o.o.
Ljubljanska c. 24a, Kranj, Slovenija

e-mail: sergej.rozman@abakus.si

phone: +386 4 287 11 14