NetFlow Data Capturing and Processing at SWITCH and ETH Zurich

Arno Wagner

wagner@tik.ee.ethz.ch

Communication Systems Laboratory

Swiss Federal Institute of Technology Zurich (ETH Zurich)

Talk Outline

- The DDoSVax Project
- The SWITCH Network
- NetFlow Data Capturing Infrastructure
- Long-Term Storage
- Computing infrastructure
- Infrastructure Cost
- Remarks and Lessons Learned
- Online Processing Framework: UPFrame
- Conclusion

The DDoSVax Project

`http://www.tik.ee.ethz.ch/~ddosvax/`

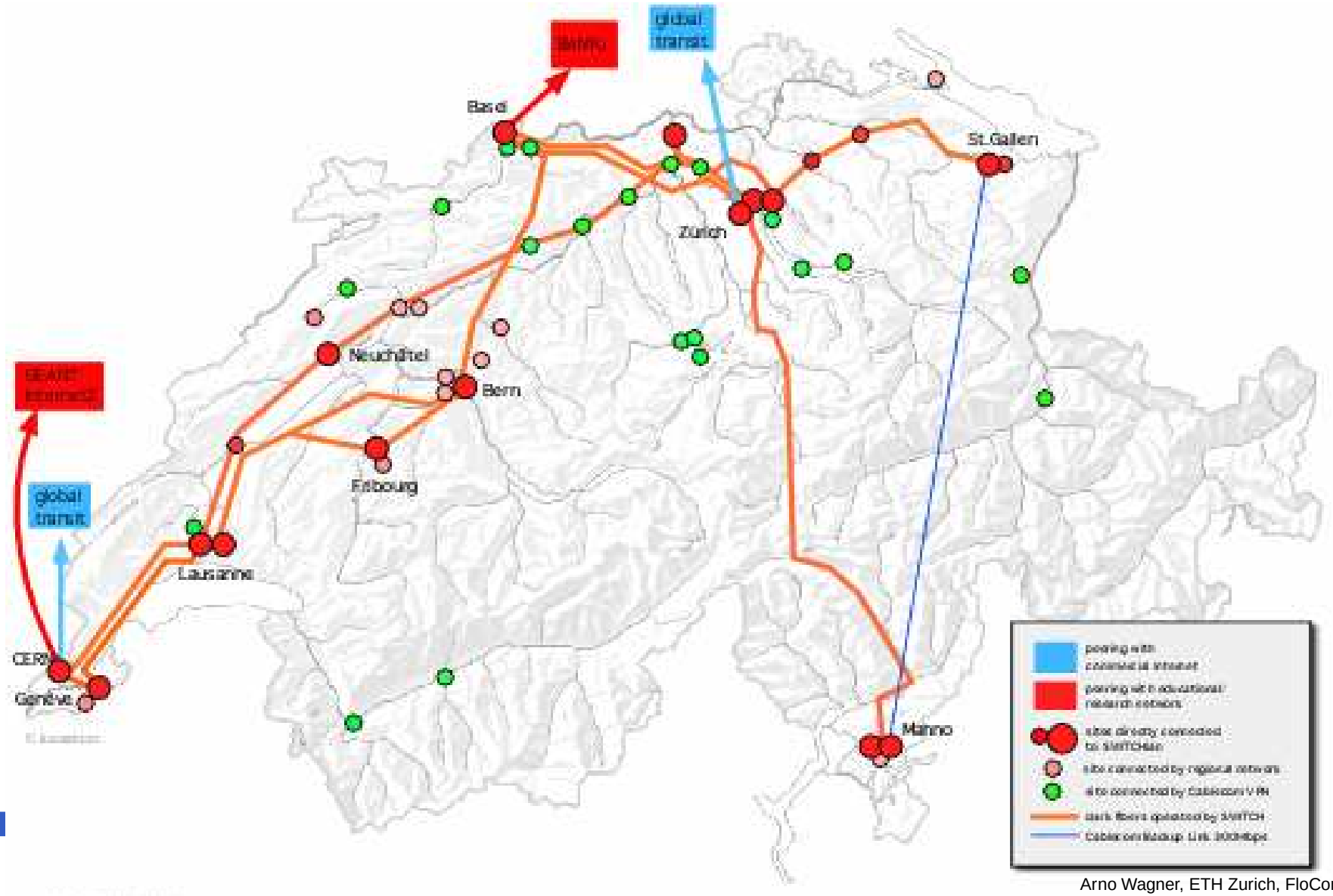
- Collaboration between SWITCH (www.switch.ch) and ETH Zurich (www.ethz.ch)
- Aim (long-term): Analysis and countermeasures for DDoS-Attacks and Internet Worms
- Start: Begin of 2003
- Funded by SWITCH and the Swiss National Science Foundation

SWITCH

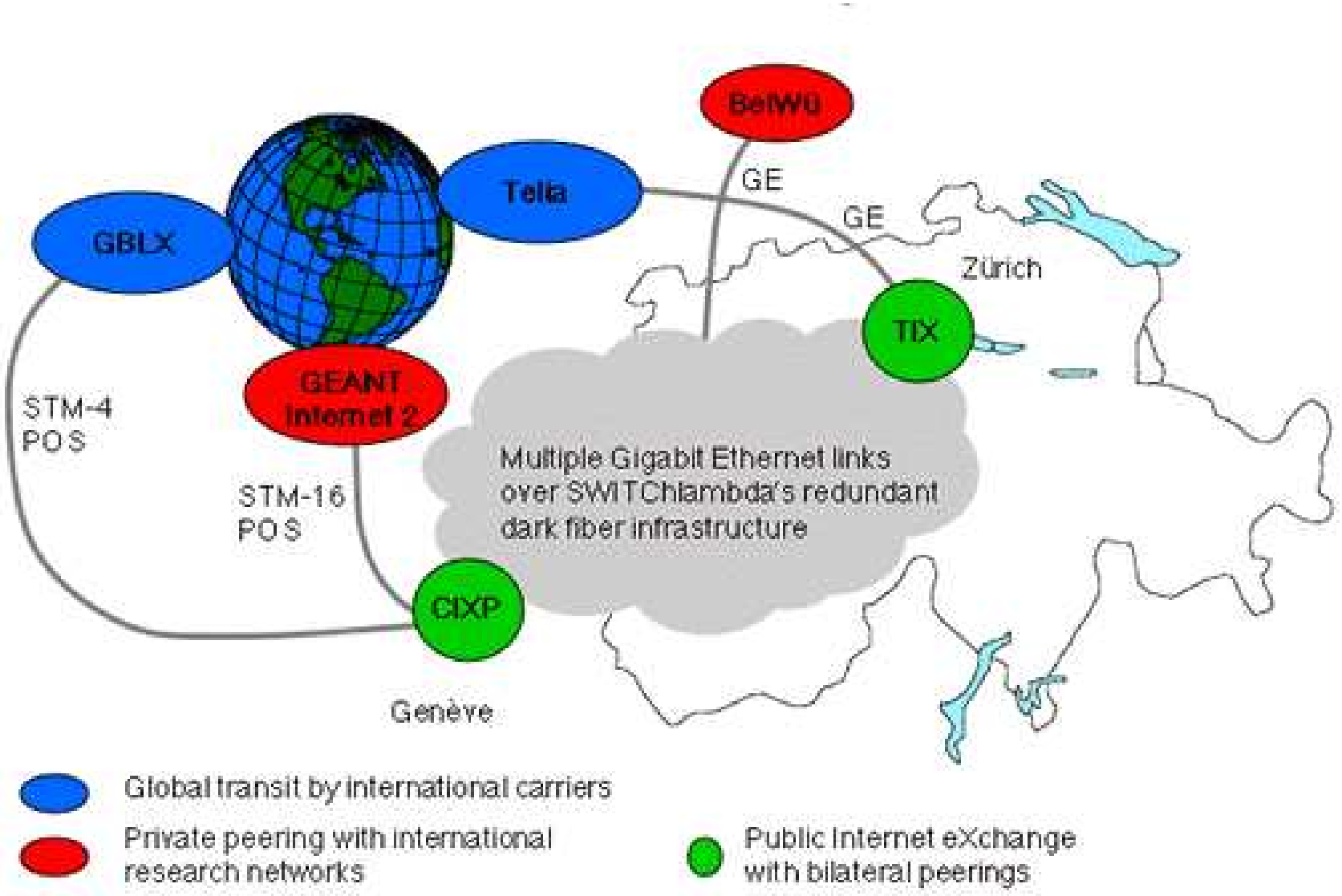
The Swiss Academic And Research Network

- .ch Registrar
- Links most (all?) Swiss Universities
- Connected to CERN
- Carried around 5% of all Swiss Internet traffic in 2003
- Around 60.000.000 flows/hour
- Around 300GB traffic/hour

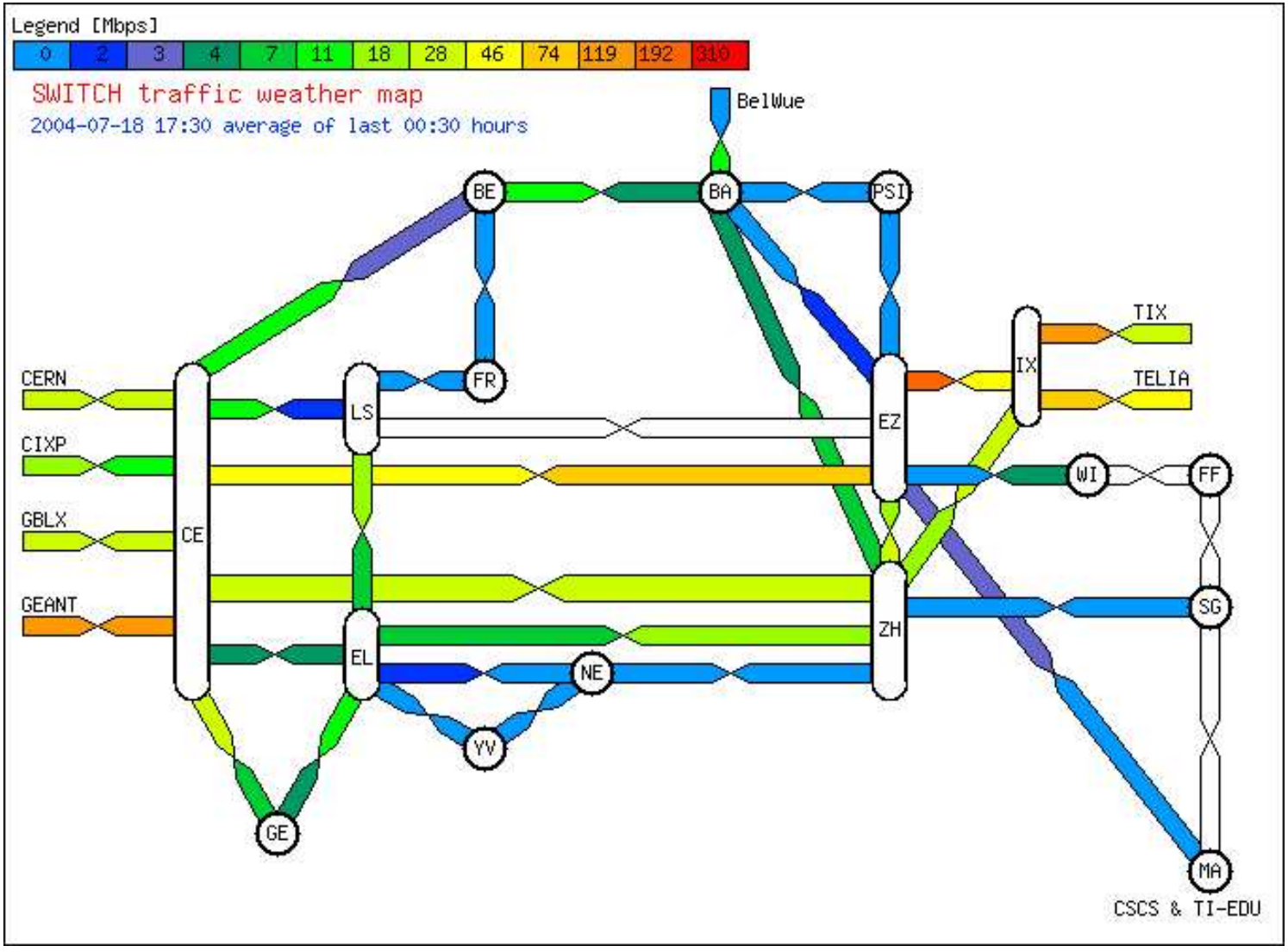
The SWITCH Network



SWITCH Peerings



SWITCH Traffic Map



SWITCH Routers

(Don’t ask me for specifics...)

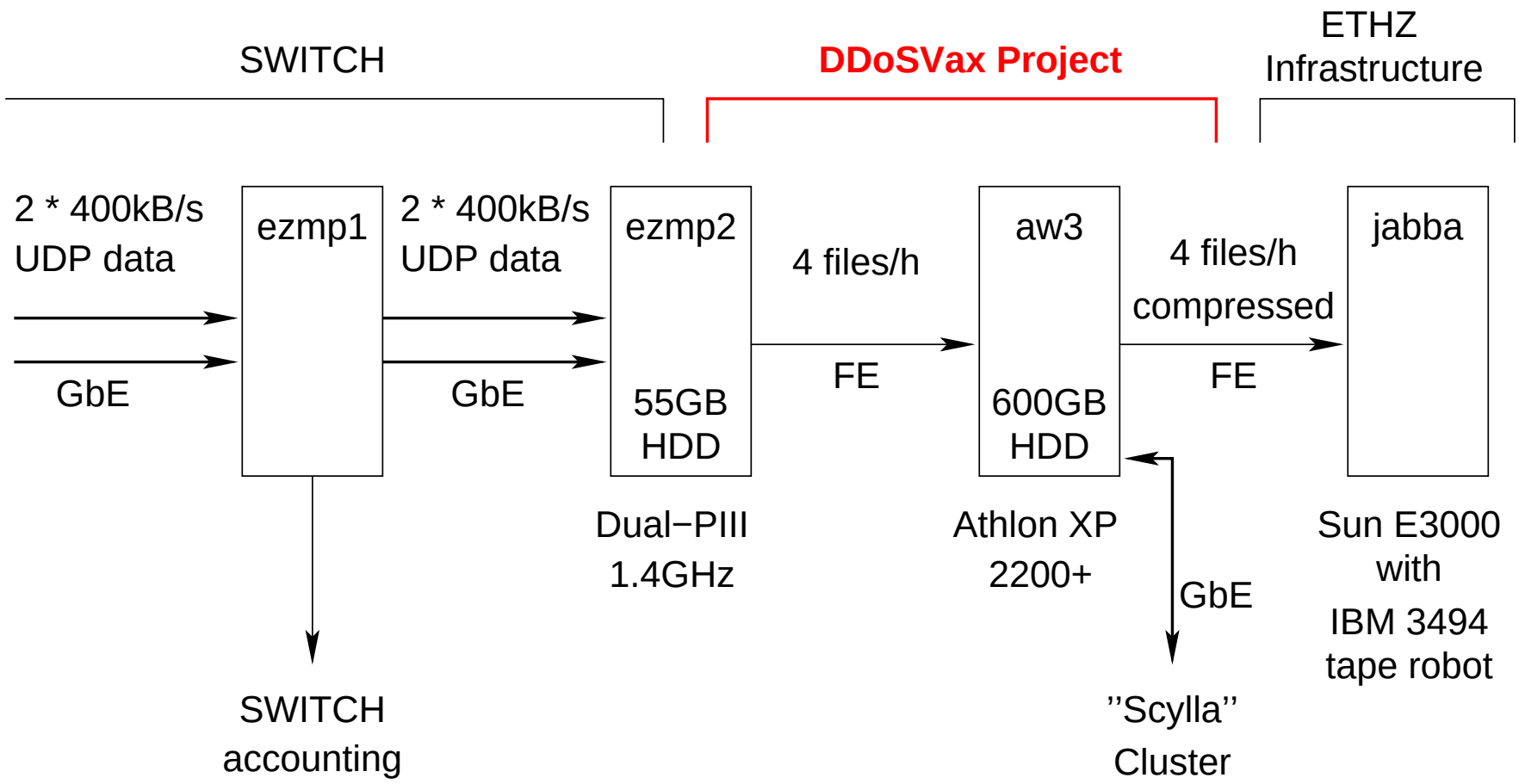
- swiCE2, swiCE3, swiIX1: Cisco 7600 OSR with Supervisor 720
- swiBA2: Cisco 7600 OSR with Supervisor 2
- Cards: 8/16/48 GbE, 10GbE
- OSM POS OC-48c
- OSM POS 2*OC-12c
- OSM 4*Gigabit Ethernet

NetFlow Data Usage at SWITCH

- Accounting
- Network load monitoring
- SWITCH-CERT, forensics
- DDoSVax (with ETH Zurich)

Transport: Over the normal network

NetFlow Data Flow



NetFlow Capturing

- One Perl-script per stream
- Data in one hour files
- Timestamps and src-IP in ”stat” file

Critical: Linux socket buffers:

- Default: 64kB/128kB max.
- Maximal possible: 16MB
- We use 2MB (app-configured)
- 32 bit Linux: May scale up to 5MB/s per stream

Capturing Redundancy

- Worker / Supervisor (both demons)
- Super-Supervisor (cron job)
  For restart on reboot or supervisor crash
- Space for 10-15 hours of data

No hardware redundancy

Data Transfer to ETHZ

- Cron job, every 2 hours
- Single Perl script
- Transfer: scp (no compression, RC4)
- Remote deletion: ssh

No compression on ezmp2. (Some other Software running there)
Bzip2 compression on ezmp2 would be possible!

Long-Term Storage Format

Full data since March 2003
Bzip2 compressed raw NetFlow V5 in one-hour files

- We need most data and precise timestamps
- We don’t know what to throw away
- We have the space
- Preprocessing for specific work still possible

Latency: 5-10 minutes / hour of data

Computing Infrastructure

The ”Scylla” Cluster
Servers:

- aw3: Athlon XP 2200+, 600GB RAID5, GbE
- aw4: Dual Athlon MP 2800+, 800GB RAID5, GbE
- aw5: Athlon XP 2800+, 800GB RAID5, GbE

Nodes:

- 22 * Athlon XP 2800+, 120GB, GbE

Infrastructure Cost Today

Speaker: 1 MYr = 175.000 CHF = 142.000 USD
$\Rightarrow$ 1MM = 12.000 USD, 1MD = 640 USD
Hardware and full installation:

- aw3 (capturing): 1600 USD + 2 MD
- aw4 (dual CPU server): 2500 USD + 3 MD
- Cluster: 24.000 USD + 1MM
- Maintenance: 1-2 MD/month

Hidden cost: Computer room, network infrastructure, software development
Scalability: Add 2*200GB HDD to each node
$\Rightarrow$ 8TB additional at 6000 USD

Lessons learned

Most important: KISS!

- Use scripting wherever possible
- Worker and Supervisor pairs are simpler
  $\Rightarrow$ ”crash” as error recovery model
- Cron as basic reliable execution service
- Email for notification: Do rate-limiting
- File-copy: Interlock and age check
- ssh, scp password-less (user key)
- Nothing needs to run as ”root”!

Remarks on Software

- Linux is stable enough
- Linux is fast enough
- Linux Software RAID1/5 works well
- XFS has issues with Software RAID
- Perl is suitable for demons
- Python is suitable for demons

Remarks on Hardware

PC hardware works well, but:

- Get good quality components (PSUs!)
- Get good cooling (HDDs/CPUs)
- Do SMART monitoring
- Do regular complete surface scans
- Have cold spares handy
- ...

Remarks on Linux Clusters

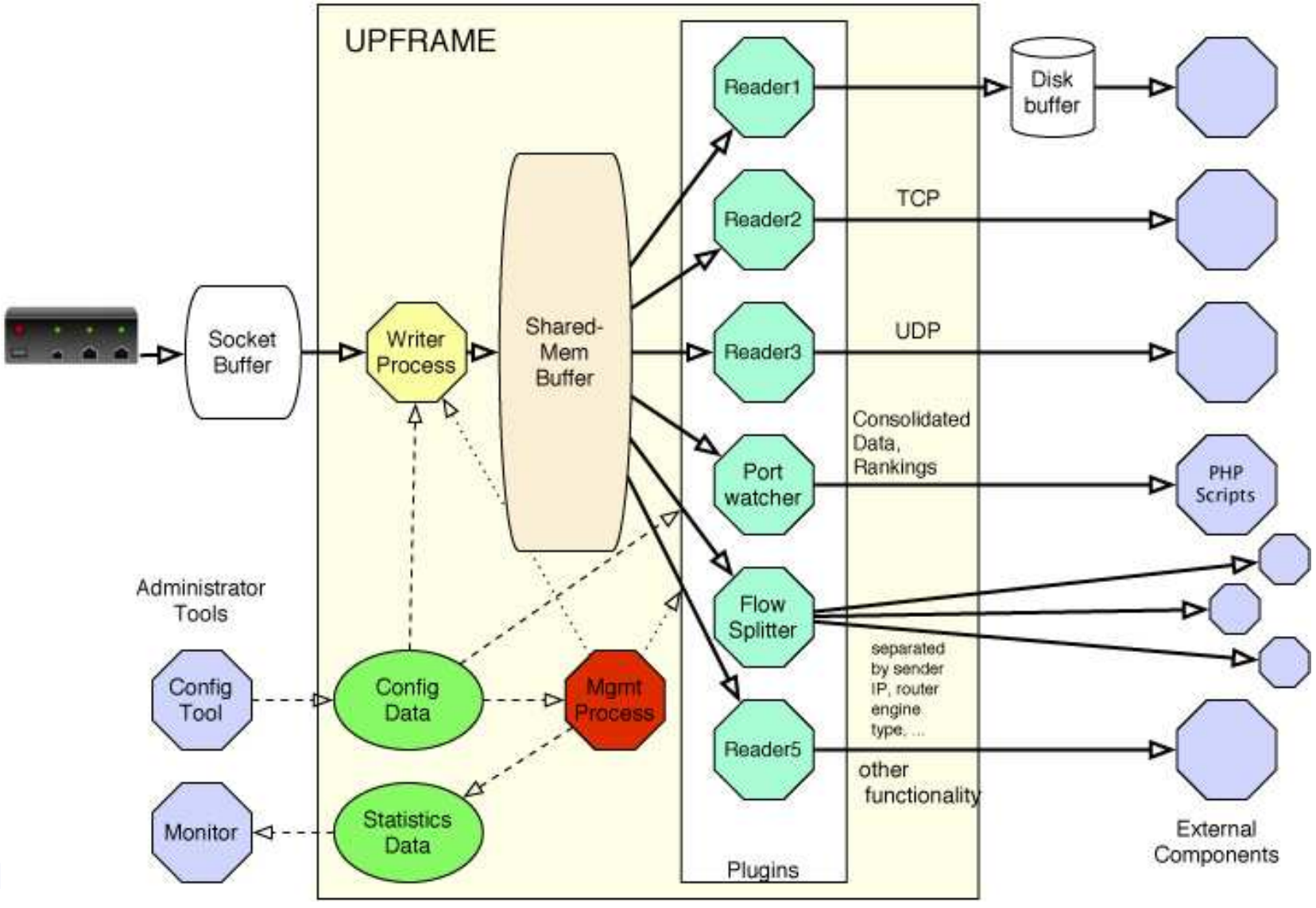
- Rackmount vs. ”normal”
- Cooling / Power needs planning
- Gigabit Ethernet "star" topology is nice
- KVM not for all nodes needed
- FAI (Fully Automatic Installation) for installation
- Local Debian mirror
  $\Rightarrow$ 10 Min for complete reinstallation
- No global connectivity for the nodes
- Private addresses for the nodes

UPFrame

`http://www.tik.ee.ethz.ch/~ddosvax/upframe/`

- UDP plugin framework
- E.g. for online analysis of NetFlow data
- Can be used as traffic-shaper
- Robust: For experimental plugins

UPFrame Structure



Conclusion

- SWITCH is large enough and small enough
- No special hardware / software needed for capturing
- Long-term storage is unproblematic
- Linux can be used in the whole infrastructure
- Online processing is more difficult
- Simplicity and Reliability are the main issues
- ...