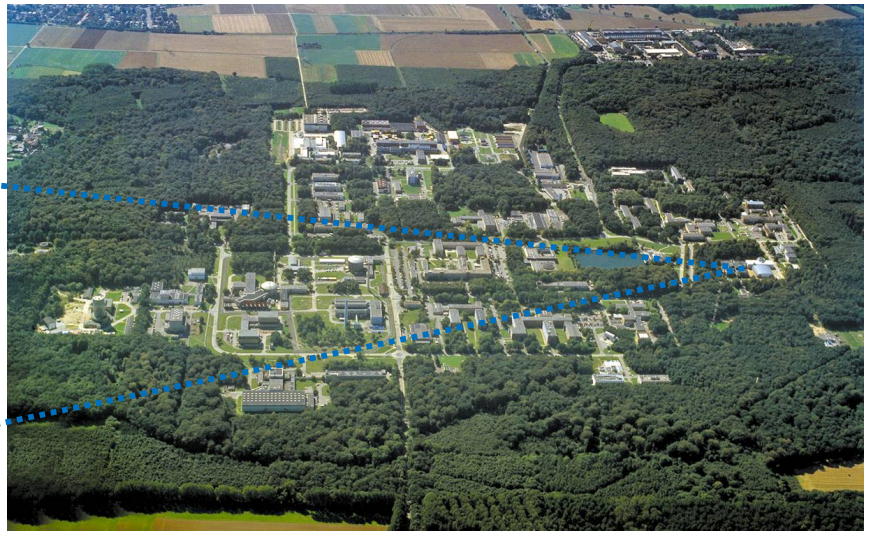
First experiences configuring a perfSONAR mesh in an PRACE MDVPN environment

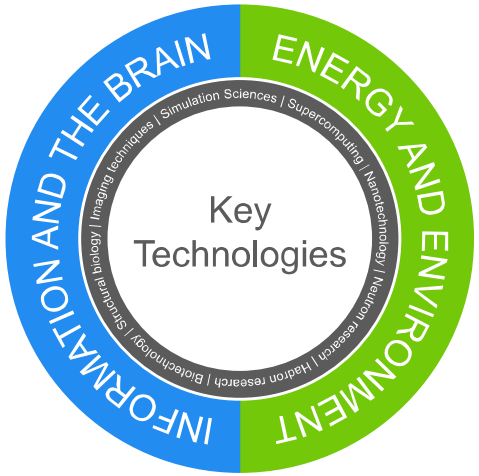
05.06.2019 *Ralph Niederberger* *rrn@fzj.de* *Forschungszentrum Jülich GmbH*

JÜLICH Forschungszentrum

Member of the Helmholtz Association

Forschungszentrum Jülich GmbH at a glance





- **Budget:** 610 Mio €,
  including 245 Mio € third party funding
  100 Horizon 2020 projects, 420 national projects
- **Employees:** 5.900
  incl. 1.950 scientists including PhD students
  800 guest scientists from 75 countries
- **Publications:** 2.450
  (source: fact sheet 2017)

Jülich Supercomputing Centre (JSC)

Facts and Figures

Staff:

220 Total (185 FTE)

160 Scientists

13 PhD Students (+13 external)

Budget:

30 Mio. € Institutional Funding (PoF)

15 Mio. € Third Party Funding

Jülich Supercomputing Centre (JSC)

- The Jülich Supercomputing Centre operates supercomputers of the highest performance class.
- It enables scientists and engineers to solve their highly complex problems by simulations.
- Currently, we are part of several EU projects like PRACE, HBP, EOSC-Hub, AENEAS and a lot of others all related to HPC or Big Data.

- So networking is one of the most relevant parts of our job.

PRACE in a few words

- The mission of PRACE (Partnership for Advanced Computing in Europe) is to enable high-impact scientific discovery and engineering research and development across all disciplines to enhance European competitiveness for the benefit of society. PRACE seeks to realize this mission by offering world class computing and data management resources and services through a peer review process.

- PRACE also seeks to strengthen the European users of HPC in industry through various initiatives.

- PRACE has a strong interest in improving energy efficiency of computing systems and reducing their environmental impact.

PRACE | members

Hosting Members

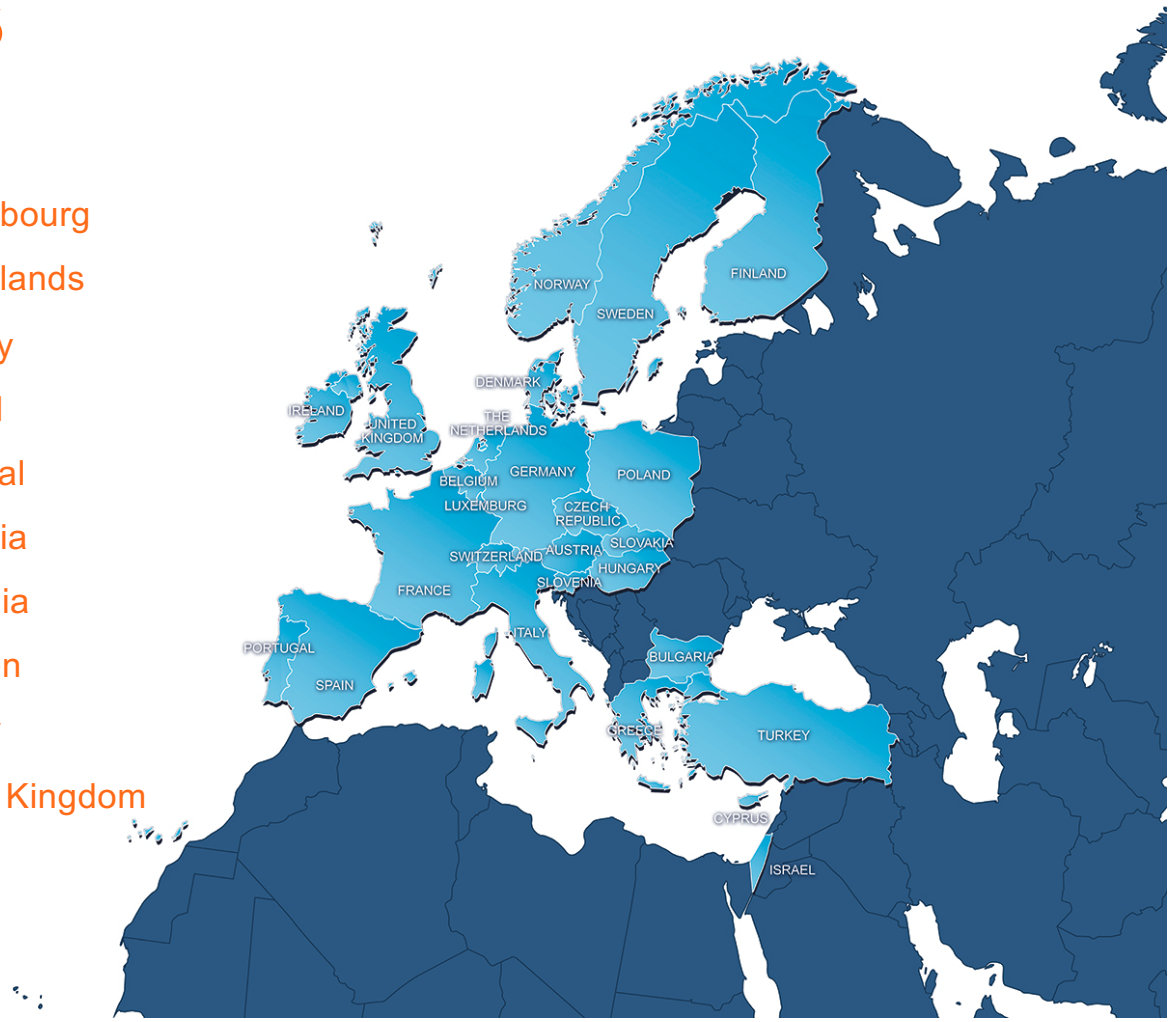
- France
- Germany
- Italy
- Spain
- Switzerland

General Partners (PRACE 2)

- Austria
- Belgium
- Bulgaria
- Cyprus
- Czech Republic
- Denmark
- Finland
- Greece
- Hungary
- Ireland
- Israel
- Luxembourg
- Netherlands
- Norway
- Poland
- Portugal
- Slovakia
- Slovenia
- Sweden
- Turkey
- United Kingdom

Observers

- Croatia
- Romania



PRACE | what we do

- Open access to world-class HPC systems to EU scientists and researchers
- Variety of architectures to support the different scientific communities
- High standards in computational science and engineering
- Peer Review at European level to foster scientific excellence
- Robust and persistent funding scheme for HPC supported by national governments and European Commission (EC)
- Support the development of intellectual property rights (IPR) in Europe by working with industry and public services
- Collaborate with European HPC industrial users and suppliers

PRACE | achievements

- 688 scientific projects enabled
- >21 000 000 000 (thousand million) core hours awarded since 2010
- Of which 63% led by another PI nationality than the HM
- R&D access to industrial users with >50 companies supported
- >12 000 people trained through PRACE Training
- ~110 Petaflops of peak performance on 7 world-class systems
- 26 PRACE members, including 5 Hosting Members
  (France, Germany, Italy, Spain and Switzerland)
- PRACE is the only e-infrastructure Landmark on the ESFRI Roadmap 2016

PRACE | Tier-0 Systems in 2018

NEW ENTRY 2018
JUWELS (Module 1): Bull Sequana
GAUSS @ FZJ, Jülich, Germany
#26 Top 500

MareNostrum: IBM
BSC, Barcelona, Spain
#25 Top 500

Piz Daint: Cray XC50
CSCS, Lugano, Switzerland
#5 Top 500

NEW ENTRY 2018/2019
SuperMUC NG : Lenovo cluster GAUSS @ LRZ, Garching, Germany #8 Top 500

NEW ENTRY 2018
JOLIOT CURIE : Bull Sequana
GENCI/CEA, Bruyères-le-Châtel, France #40 Top 500

MARCONI: Lenovo
CINECA, Bologna, Italy
#19 Top 500

Hazel Hen: Cray
GAUSS/HLRS, Stuttgart, Germany
#30 Top 500

Close to 110 Petaflops cumulated peak performance

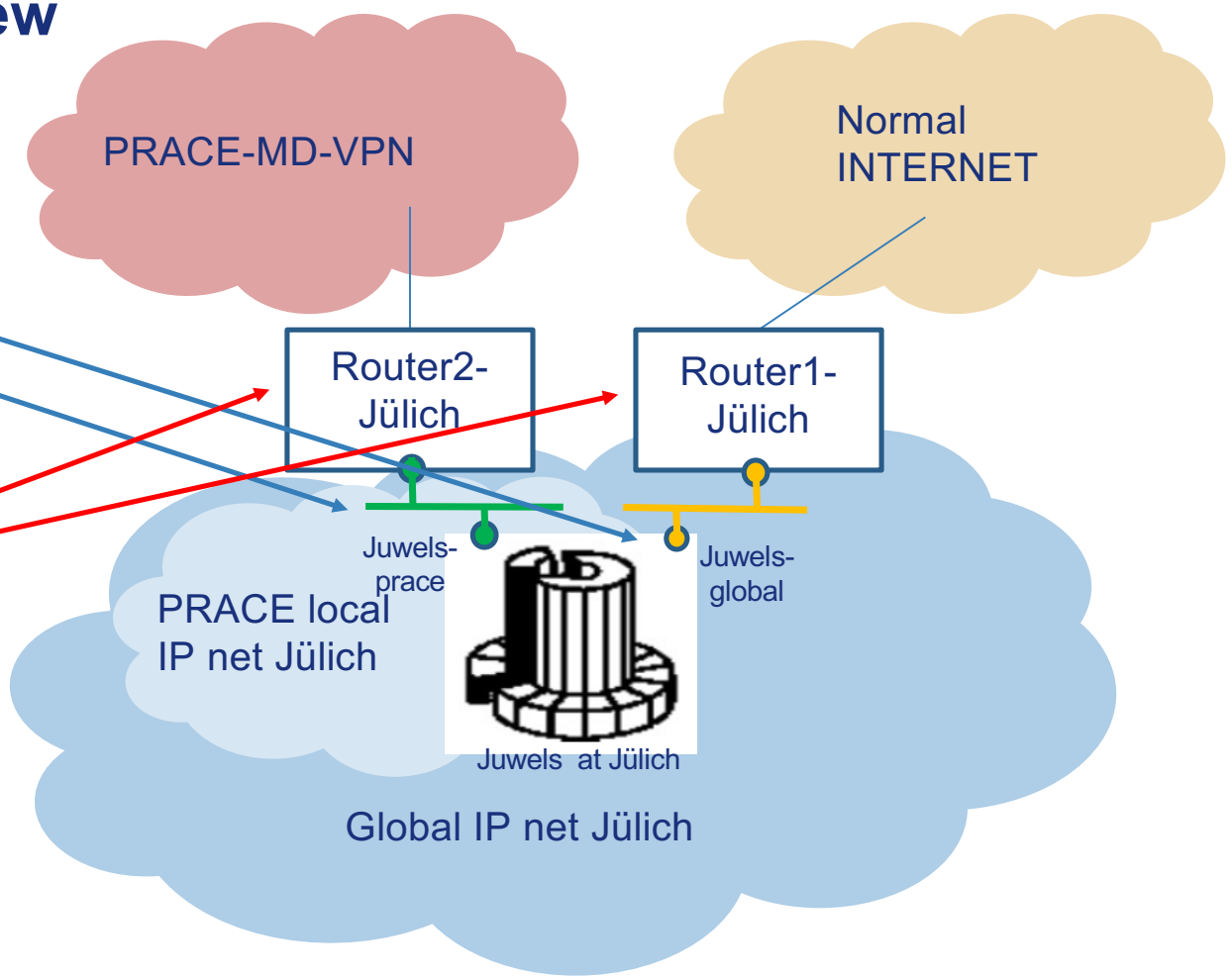
Getting things together

- Prace partners are connected to each other via a MD-VPN provided by GÉANT allowing fast access between HPC systems.
- Firewalls may be implemented in between, but a „Net of Trust“ idea doesn‘t necessitate this.
- But what about bandwidth?
- Achieving optimal end-to-end performance is a multi-faceted problem including:
  - Appropriate network capacity provisioning between the end sites
  - Properties of the local campus network (at each end), including capacity of the external connectivity, internal LAN design, the performance of firewall / IDS devices, and the configuration of other devices on the path
  - End system configuration and tuning; network stack buffer sizes, disk I/O, …
  - The choice of tools used to transfer data, e.g. scp, Globus, rsync, Aspera, …
- To optimise end-to-end performance, you need to address each aspect
- Nevertheless, there will inevitably be a bottleneck somewhere

PRACE MDVPN Network overview

The HPC system‘s view

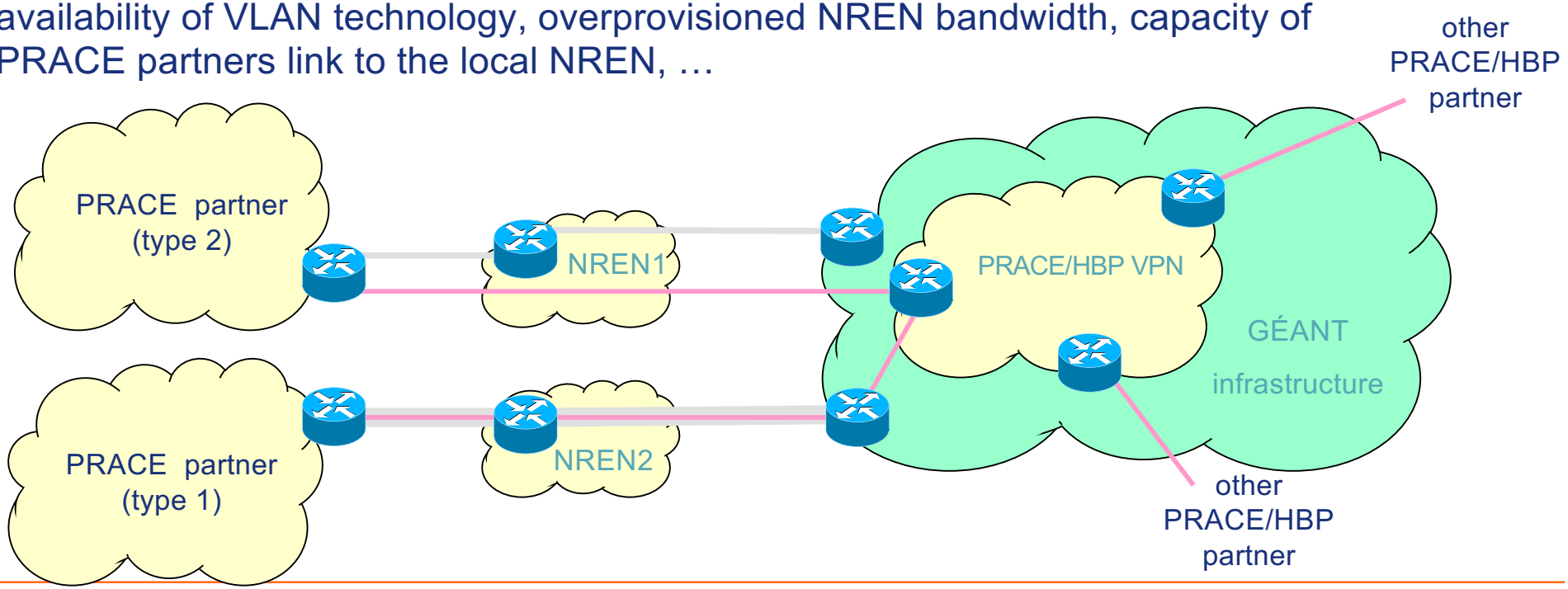
- The two links can be dedicated or shared
- As well the routers may be dedicated ones or virtual routers



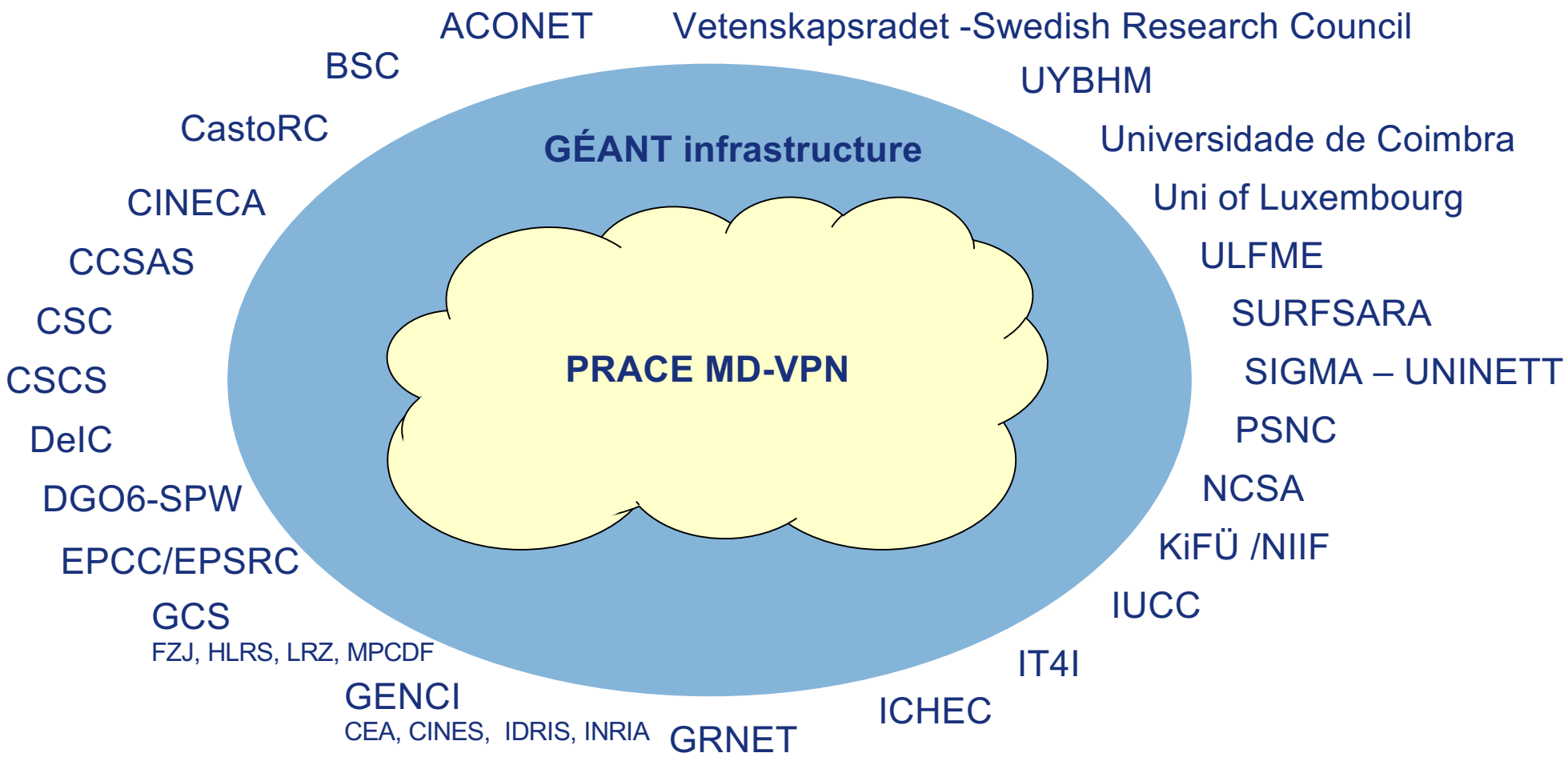
Types of connectivity

Links from partner sites to the PRACE VPN on GÉANT infrastructure to the SDP can be implemented

- as VLAN on the existing site links via their NRENs to GÈANT (type 1) or
- as dedicated links (type 2) dependend on the potential of the local NRENs, i.e. availability of VLAN technology, overprovisioned NREN bandwidth, capacity of PRACE partners link to the local NREN, …



The logical PRACE network view



Why perfSONAR and why a mesh configuration

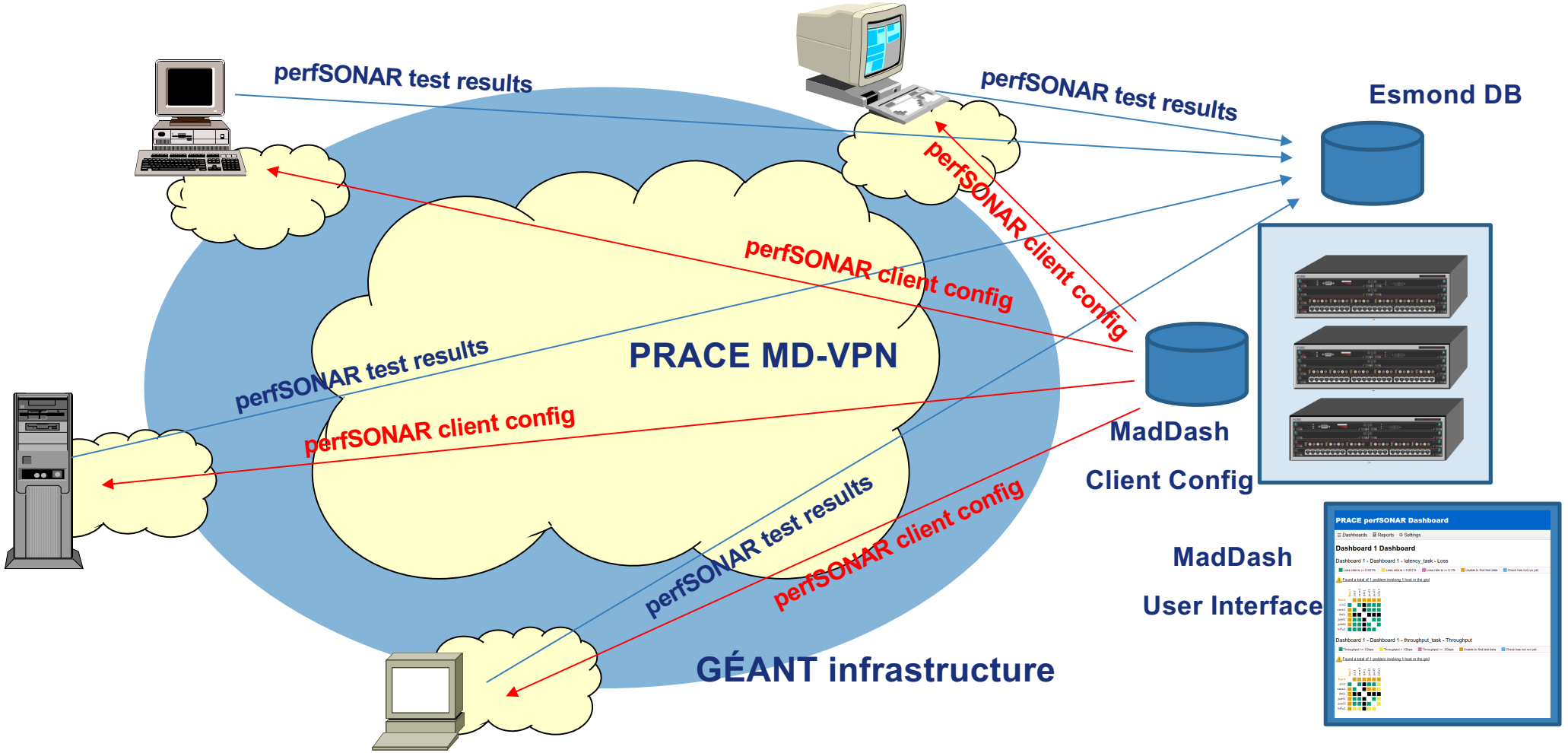
- PRACE has a network monitoring system based on iperf and a self-developed client-server infrastructure since the beginning.
- Over the years systems came and left, including admins, so that adaptations to the software and education of personal had been necessary, often again and again.
- Furthermore long running iperf servers led to measurement outages, because of undefined server stati (hanging).
- Configuration of iperf servers, cronjobs, checking of logs, etc. time consuming.
- Admins not well prepared for network optimizations, since network personal not „PRACE“ related.
- So a network tool, independent of HPC admin work and only network related, would help a lot.

PRACE perfSONAR usage – the original idea

- PerfSONAR systems at any location with adequate interface connection,

  i.e. similar to local HPC system
- Advantages:
  - Independent of HPC system
  - Optimized configuration
  - No influence on performance of/on HPC system
- Disadvantages:
  - Further system needed (costs, interfaces, administration, security)

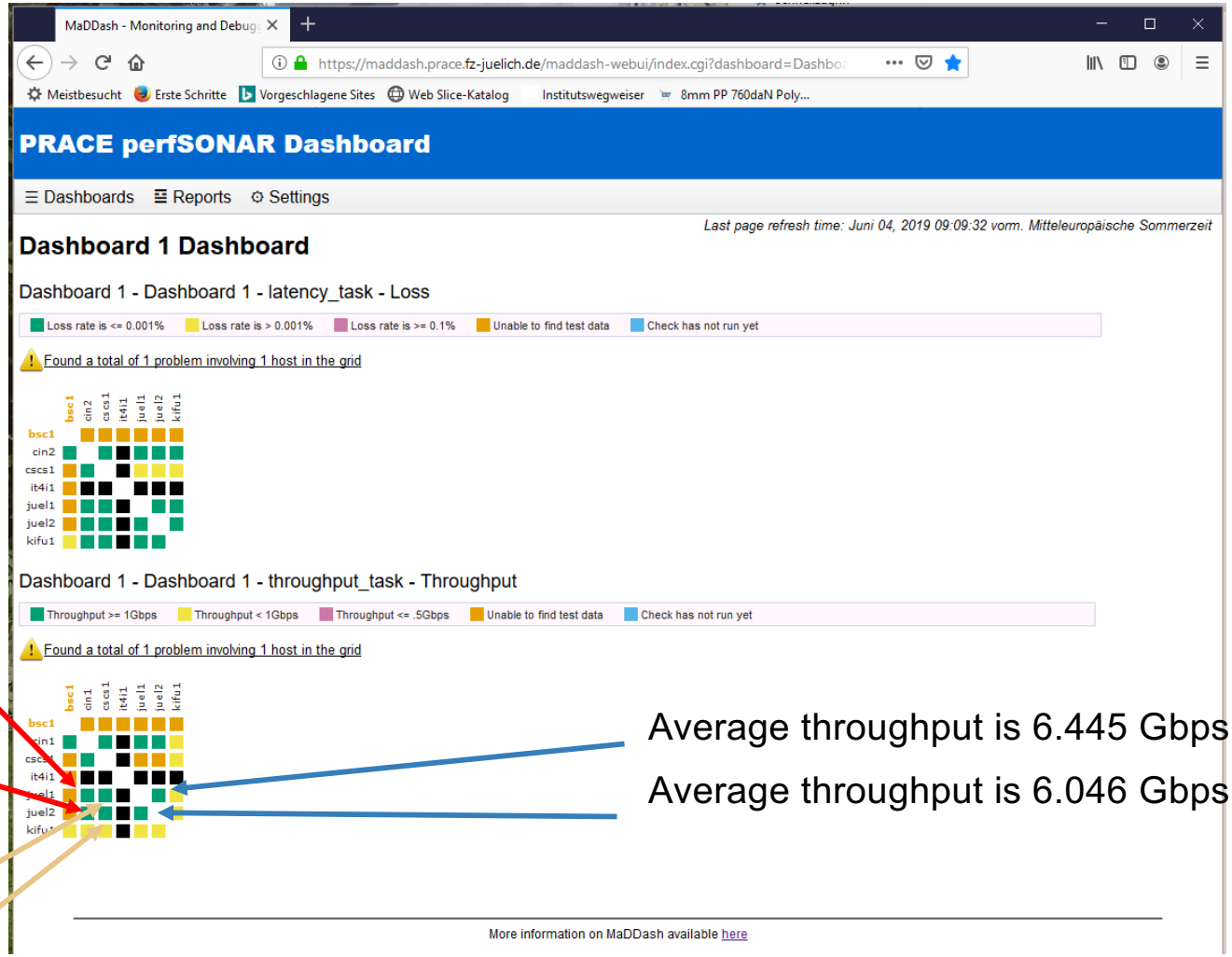
The PRACE perfSONAR mesh



PRACE perfSONAR status

- Several systems have been installed at different partner sites
- E.g. in Jülich several systems have been prepared
  a) An old standalone 19'' rack system with 10 Gb/s interface card
  b) the standard judac server system (DTN) (100 Gb/s interface card)
  c) A Maddash server system for
     a) *controlling test schedules and*
     b) *collecting test results as well as*
     c) *a web frontend system for presenting results to PRACE users*
- Other partners installed dedicated systems or virtual systems with dedicated or shared interfaces

The current PRACE perfSONAR Dashboard



Average throughput is 2.665 Gbps

Average throughput is 1.718 Gbps

Average throughput is 6.445 Gbps

Average throughput is 6.046 Gbps

Average throughput is 3.942 Gbps

Average throughput is 1.184 Gbps

☰ Dashboards ☰ Reports ⚙ Settings

Last page refresh time: Juni 04, 2019 14:38:41 nachm. Mitteleuropäische Sommerzeit

juel1 to kifu1 (Throughput)

Queries an esmond MA for throughput data and alerts on response

Status: WARNING **Last Checked:** Juni 04, 2019 12:05:27 nachm. Mitteleuropäische Sommerzeit **Next Check:** Juni 04, 2019 16:05:27 nachm. Mitteleuropäische Sommerzeit

Summary | History | Check Details

▾ Current Results

Current Status: WARNING
Result of last check: WARNING
Message For Current Status: Average throughput is 0.892Gbps
Reports: No reports found for this check
Events:

Name	Description	Start	End	Check Down
No events currently scheduled.				

▸ Statistics

▾ Graph

perfSONAR test results - documentation Share/open in new window

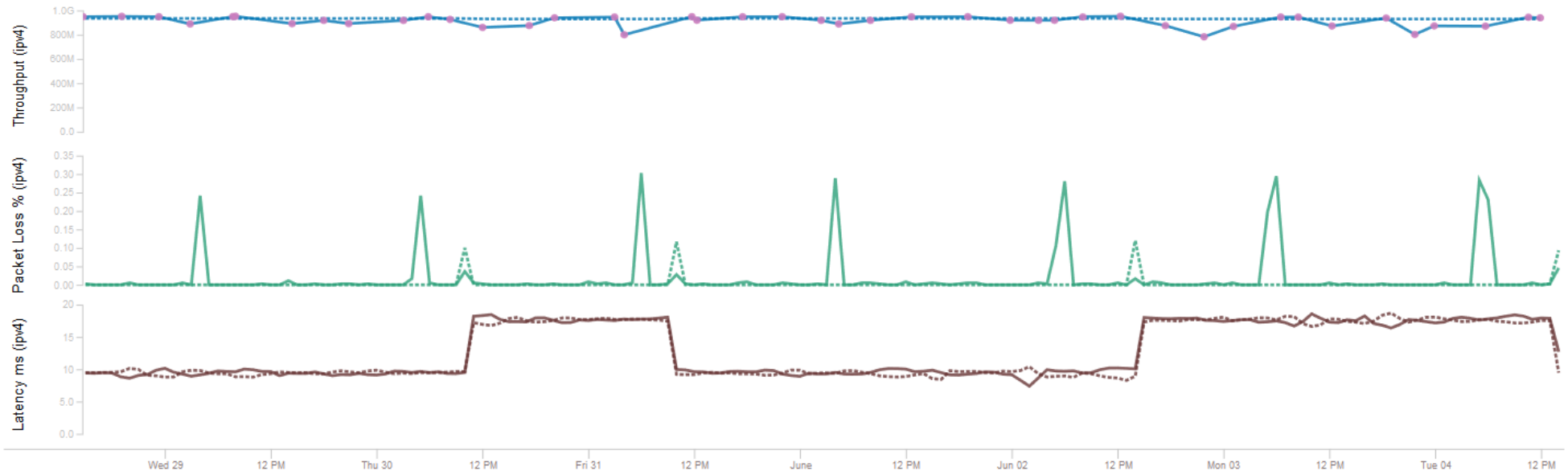
Source
perfsonar-prace.fz-juelich.de
134.94.115.220
Host info

Destination
perfsonar-prace.vh.hbone.hu
193.224.66.201
Host info

Report range
1 week
Tue 05/28/2019 14:38:42 (GMT+2) to Tue 06/04/2019 14:38:42 (GMT+2)

Tput (TCP) | Tput (UDP) | Loss (UDP) | Loss (one way) | Loss (rtt) | Retrans | Latency (one way) | Latency (rtt)

Forward | Reverse | Failures



Experiences with perfSONAR

- Testpoint Installation has been very easy
- Config from maddash central server
- It is not clear, when new mesh-pschedul.json files are really made active.
- Logging tells:
  - "… msg=Configuration file change detected, refreshing records.",
    but sometimes nothing happens.
- We saw, that
  - new configs worked directly, or
  - needed server restart, or
  - really needed power break
- With one system we had a lot of „NON-STARTER" tasks. New tasks got always deleted after 1 hour.
  Power off was needed

Experiences with perfSONAR

- Esmond and MadDash server installation straight forward
- Configuration of mesh needs a lot of document reading
- Cooking recipes for „standard“ installations would help a lot. E.g.
  - how does a standard mesh-psched.json file look like
  - Which processes have to be started
  - Where to look for error messages, if
    - *tasks don‘t run,*
    - *clients don‘t log,*
    - *results are not stored in esmond,*
    - *results are not displayed in MadDash*
  - Overall, logging is very detailed, but often confusing and required info is missing

Summary of experiences

- A good idea to start user group meetings
- Cookbooks for „standard“ installations would be very helpful.
- Stand-alone testpoints work very well (when command line tools are used)
- Establishing more complex installations cannot be done on the fly.
  - Needs a lot of reading, configuring and managing
  - Needs hand-in-hand collaborative work between test point and maddash admins
- Will provide its full power not for free.
- A lot of insight view is needed, because of complexity of interacting services.

THANK YOU FOR YOUR ATTENTION

Questions?