FOM

pdp

Showing Real Big Data

Towards visualisation of data transfers with 'Big Data' analytics techniques

HvA induction session september 2015

David Groep, Nikhef

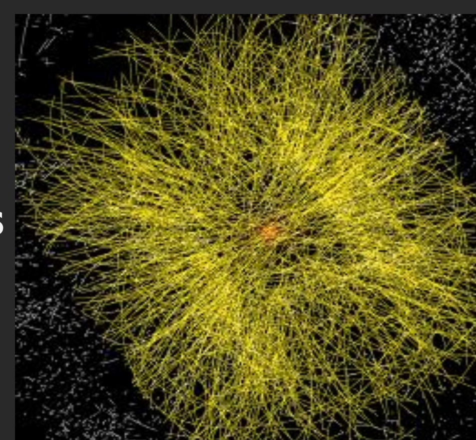
Nikhef

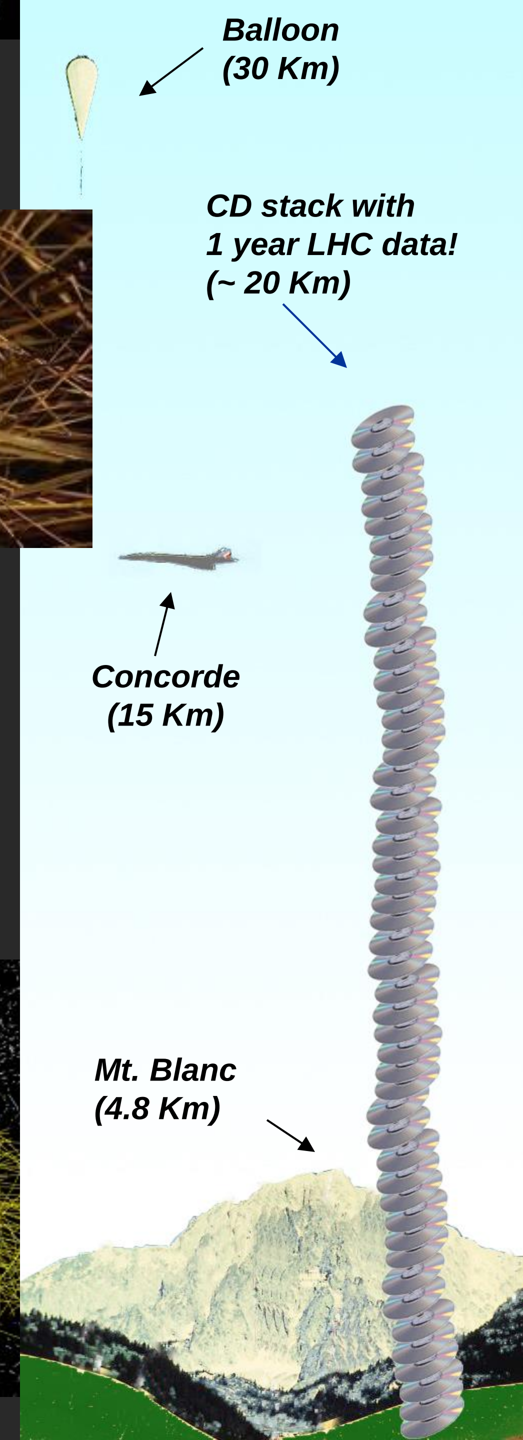
David Groep
Nikhef
Amsterdam
PDP programme

Nikhef

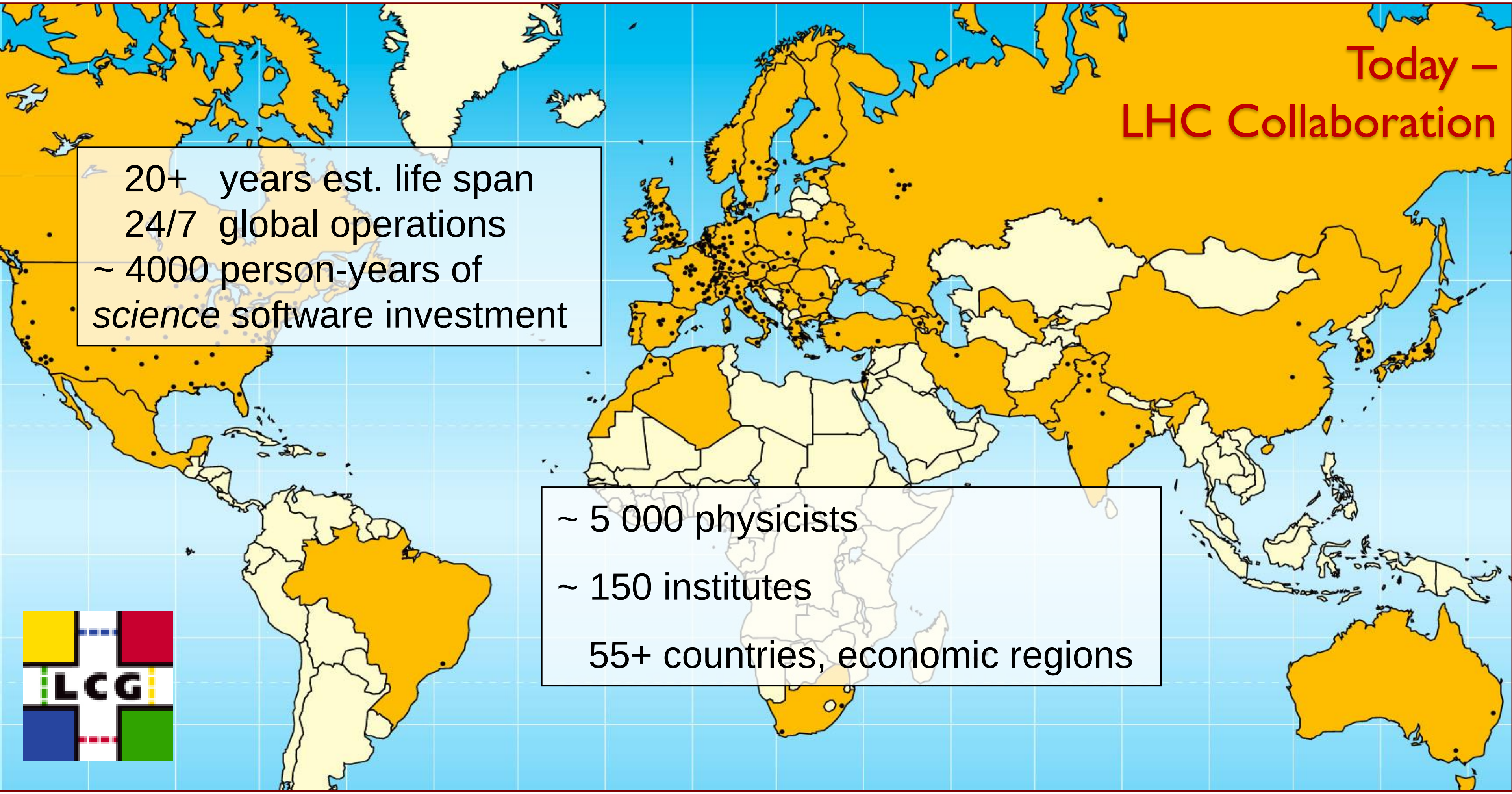
- Signal/Background 10^{-9}

- Data volume
  - (high rate) **X**
    (large number of channels) **X**
    (4 experiments)
  - **➔ 30+ PetaBytes of new data each year**
- Compute power
  - (event complexity) **X**
    (number of events) **X**
    (thousands of users)
  - **➔ 60’000 of (today's) fastest CPUs**





Today – LHC Collaboration

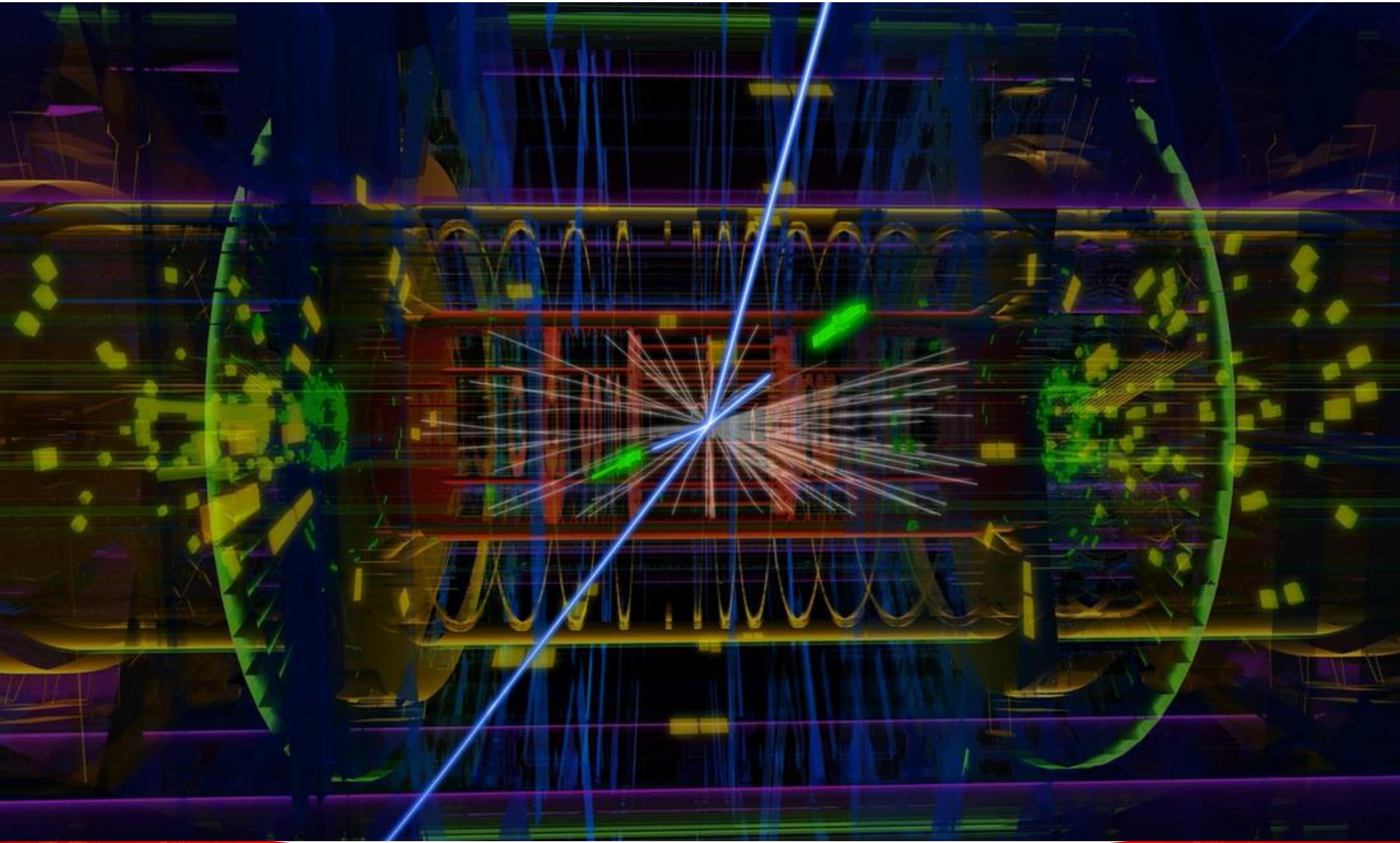


20+ years est. life span
24/7 global operations
~ 4000 person-years of *science* software investment

~ 5 000 physicists
~ 150 institutes
55+ countries, economic regions

David Groep
Nikhef
Amsterdam
PDP programme

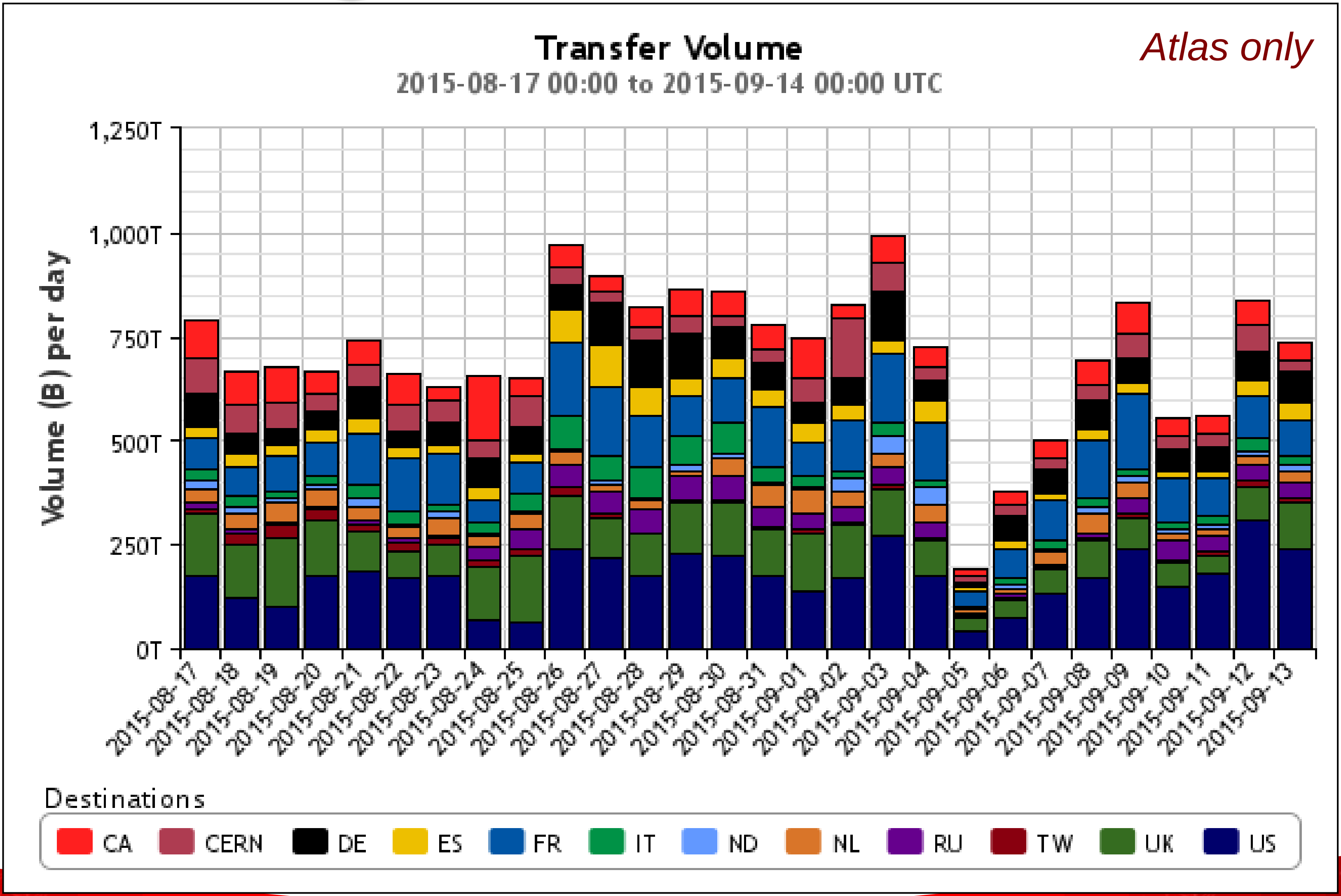
Atlas: ~50 TByte/day raw data to tape; 1000 TByte/day processed data transfers



Nikhef

Image source: CERN, Atlas

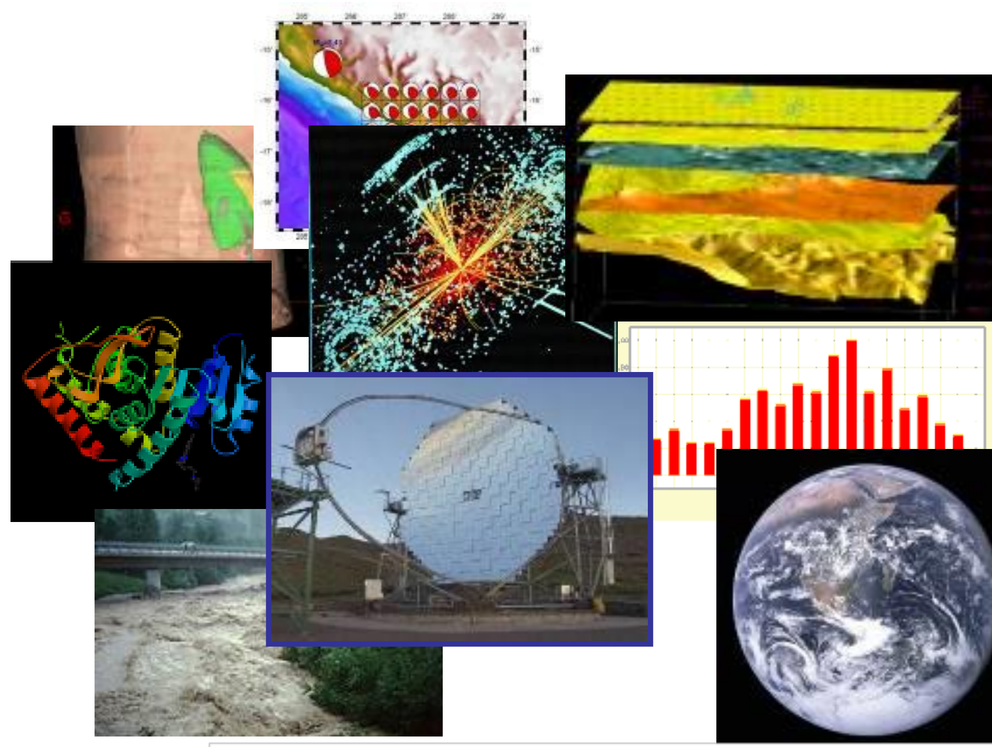
Big 'as in Large' Data

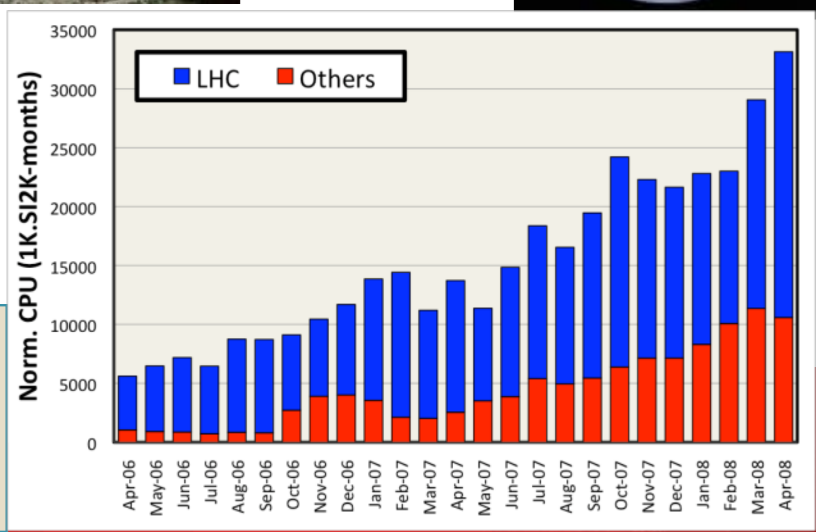


David Groep
Nikhef
Amsterdam
PDP programme

Shared e-Infrastructure

- >270 communities from many different domains
  - Astronomy & Astrophysics
  - Civil Protection
  - Computational Chemistry
  - Comp. Fluid Dynamics
  - Computer Science/Tools
  - Condensed Matter Physics
  - Earth Sciences
  - Fusion
  - High Energy Physics
  - Life Sciences
  - …





David Groep
Nikhef
Amsterdam

Applications have moved from testing to routine and daily usage

~80-95% efficiency

Global data flows

~150GByte, 12hrs
per (human) genome
per sequencer

But 1000+ sequencers…

50TByte from Shenzhen to NL is (still) done by rucksack

David Groep
Nikhef
Amsterdam
PDP programme

Genome sequencing at the Beijing Genomics Institute BGI
Photo: Scotted400, CC-BY-3.0

Distributed analysis – ‘Atlas neighbours’

0 % 100 %

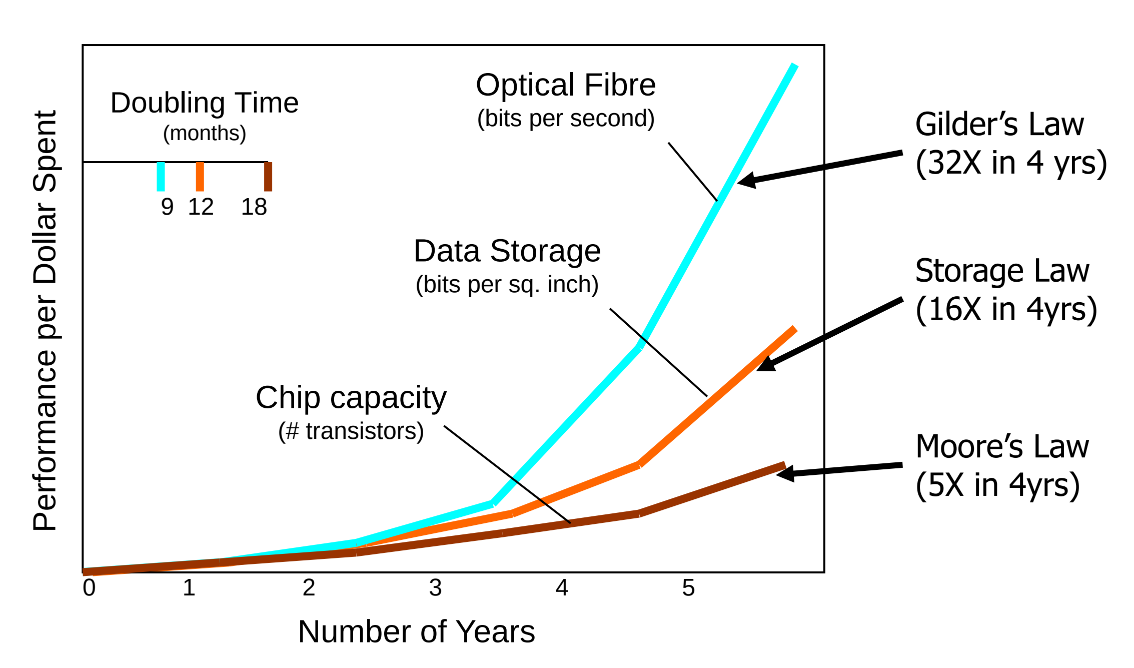
SOURCES

Displaying 12 of 12 sources and 12 of 12

	TRANSFER-	STAGING-	DELETION-	CA+	CERN+	DE+	ES+	FR+	IT+	ND+	NL+	RU+	TW+	UK+	US+
TOTAL-	97 % 98 MB/s	97 % 27 MB/s	100 % 256 MB/s	92 % 6 MB/s	95 % 726 kB/s	99 % 4 MB/s	100 % 8 MB/s	100 % 13 MB/s	99 % 6 MB/s	100 % 833 kB/s	100 % 9 MB/s	99 % 8 MB/s	100 % 12 MB/s	96 % 14 MB/s	93 % 18 MB/s
NL AM-04-YERPHI+	100 % 0 kB/s	100 % 0 kB/s	100 % 15 kB/s												
NL IL-TAU-HEP+	92 % 12 MB/s	0 % 0 kB/s	100 % 14 MB/s	100 % 3 MB/s	100 % 3 kB/s	95 % 29 kB/s	100 % 2 kB/s	100 % 2 MB/s	100 % 3 MB/s	100 % 1 kB/s	96 % 279 kB/s	90 % 301 kB/s	97 % 3 MB/s	97 % 939 kB/s	8 % 2 kB/s
NL ITEP+	92 % 1 MB/s	100 % 0 kB/s	99 % 11 MB/s	61 % 2 kB/s		100 % 1 kB/s	100 % 2 kB/s	100 % 2 kB/s		100 % 1 kB/s	100 % 1 MB/s		100 % 2 kB/s		100 % 2 kB/s
NL JINR-LCG2+	99 % 15 MB/s	100 % 0 kB/s	100 % 11 MB/s	93 % 3 kB/s	100 % 330 kB/s	100 % 313 kB/s	100 % 182 kB/s	100 % 601 kB/s	100 % 686 kB/s	100 % 764 kB/s	100 % 248 kB/s	100 % 6 kB/s	100 % 5 kB/s	100 % 2 MB/s	100 % 10 MB/s
NL NIKHEF-ELPROD+	95 % 29 MB/s	25 % 57 kB/s	100 % 26 MB/s	100 % 92 kB/s	87 % 77 kB/s	99 % 2 MB/s	100 % 7 MB/s	100 % 3 MB/s	100 % 2 MB/s	100 % 36 kB/s	100 % 3 MB/s	100 % 2 MB/s	100 % 6 MB/s	100 % 1 MB/s	94 % 4 MB/s
NL RRC-KI+	97 % 3 MB/s	0 % 0 kB/s	100 % 19 MB/s	100 % 307 kB/s		100 % 0 kB/s	0 % 0 kB/s	98 % 3 kB/s	50 % 0 kB/s	100 % 1 kB/s	100 % 1 MB/s	100 % 218 kB/s	100 % 2 kB/s	33 % 531 kB/s	67 % 454 kB/s
NL RU-MOSCOW-FIAN-LCG2+	56 % 1 MB/s	100 % 0 kB/s	100 % 11 MB/s	0 % 0 kB/s							100 % 1 MB/s	100 % 14 kB/s			
NL RU-PNPI+	96 % 621 kB/s	100 % 0 kB/s	100 % 11 MB/s	86 % 4 kB/s		100 % 1 kB/s	100 % 2 kB/s	100 % 2 kB/s	100 % 0 kB/s	100 % 1 kB/s	100 % 607 kB/s		100 % 2 kB/s		100 % 2 kB/s
NL SARA-MATRIX+	100 % 34 MB/s	99 % 27 MB/s	100 % 122 MB/s	100 % 3 MB/s	100 % 307 kB/s	100 % 2 MB/s	100 % 1 MB/s	99 % 8 MB/s		100 % 24 kB/s	100 % 2 MB/s	100 % 5 MB/s	100 % 3 MB/s	96 % 9 MB/s	98 % 1 MB/s
NL TECHNION-HEP+	100 % 51 kB/s	100 % 0 kB/s	100 % 11 MB/s	100 % 4 kB/s	100 % 9 kB/s	100 % 6 kB/s	100 % 2 kB/s	100 % 2 kB/s	100 % 0 kB/s	100 % 1 kB/s	100 % 14 kB/s	100 % 1 kB/s	100 % 2 kB/s		100 % 8 kB/s

http://dashb-atlas-ddm.cern.ch/ddm2

There's always a network close to you



SURFnet pioneered 'lambda' and hybrid networks in the world

- and likely contributed to the creation of a market for 'dark fibre' in the Netherlands

There's always fibre within 2 miles from you – where ever you are!

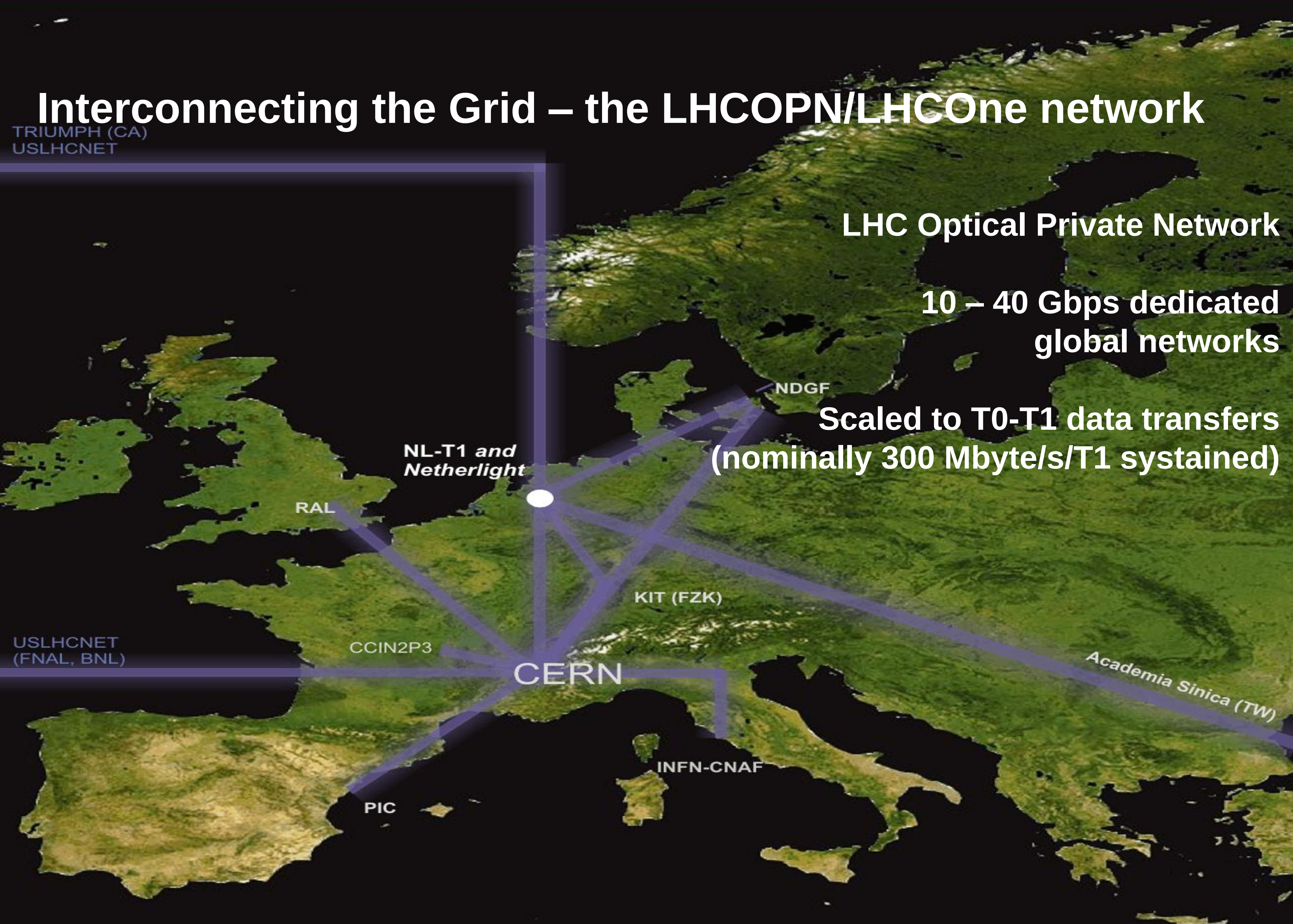
(it's just that last mile to your home that's missing – and the business model of your telecom provider…)

Interconnecting the Grid – the LHCOPN/LHCOne network

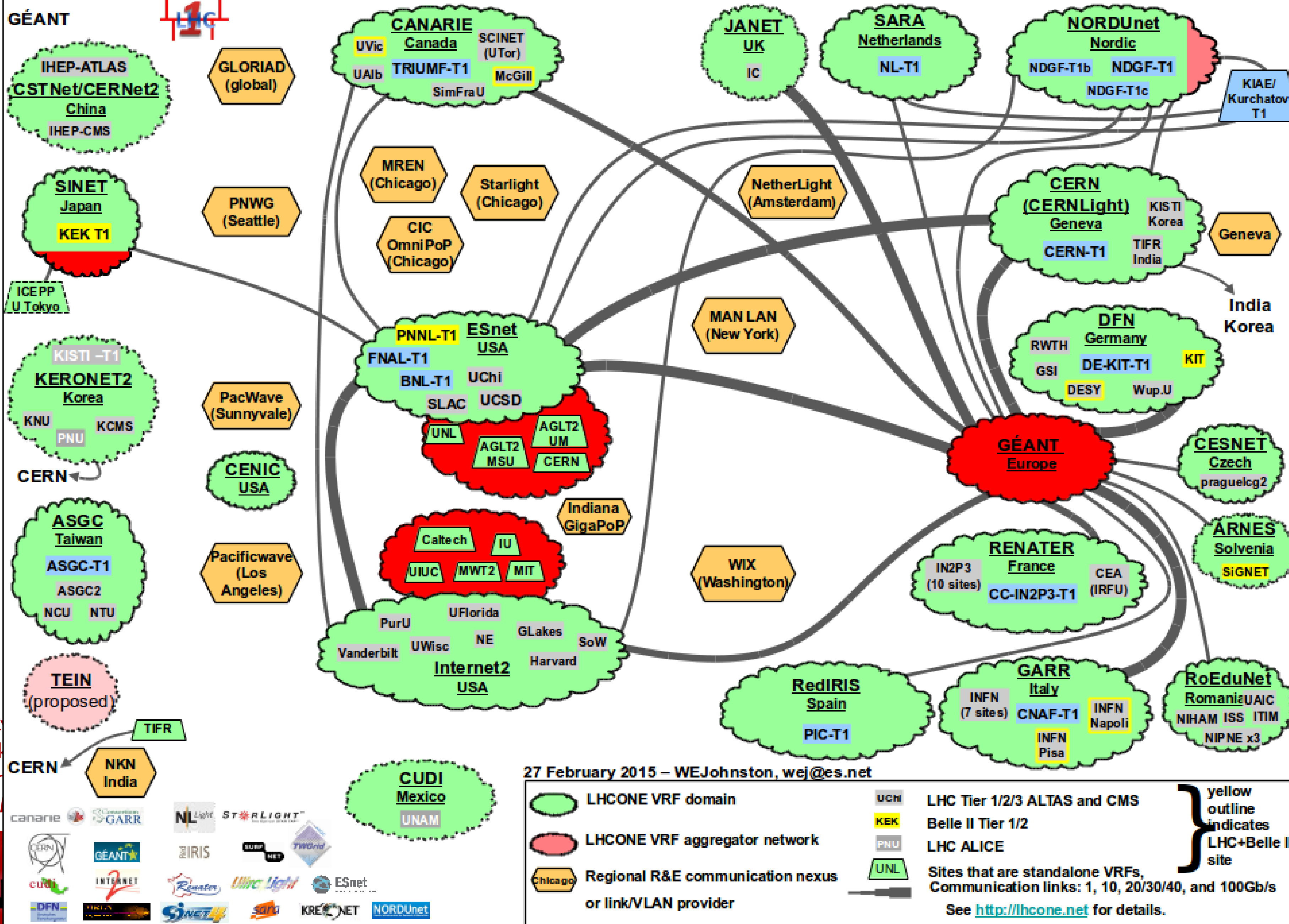
LHC Optical Private Network

10 – 40 Gbps dedicated global networks

Scaled to T0-T1 data transfers (nominally 300 Mbyte/s/T1 systained)



LHCONE: A global infrastructure for the High Energy Physics (LHC and Belle II) data management



Da

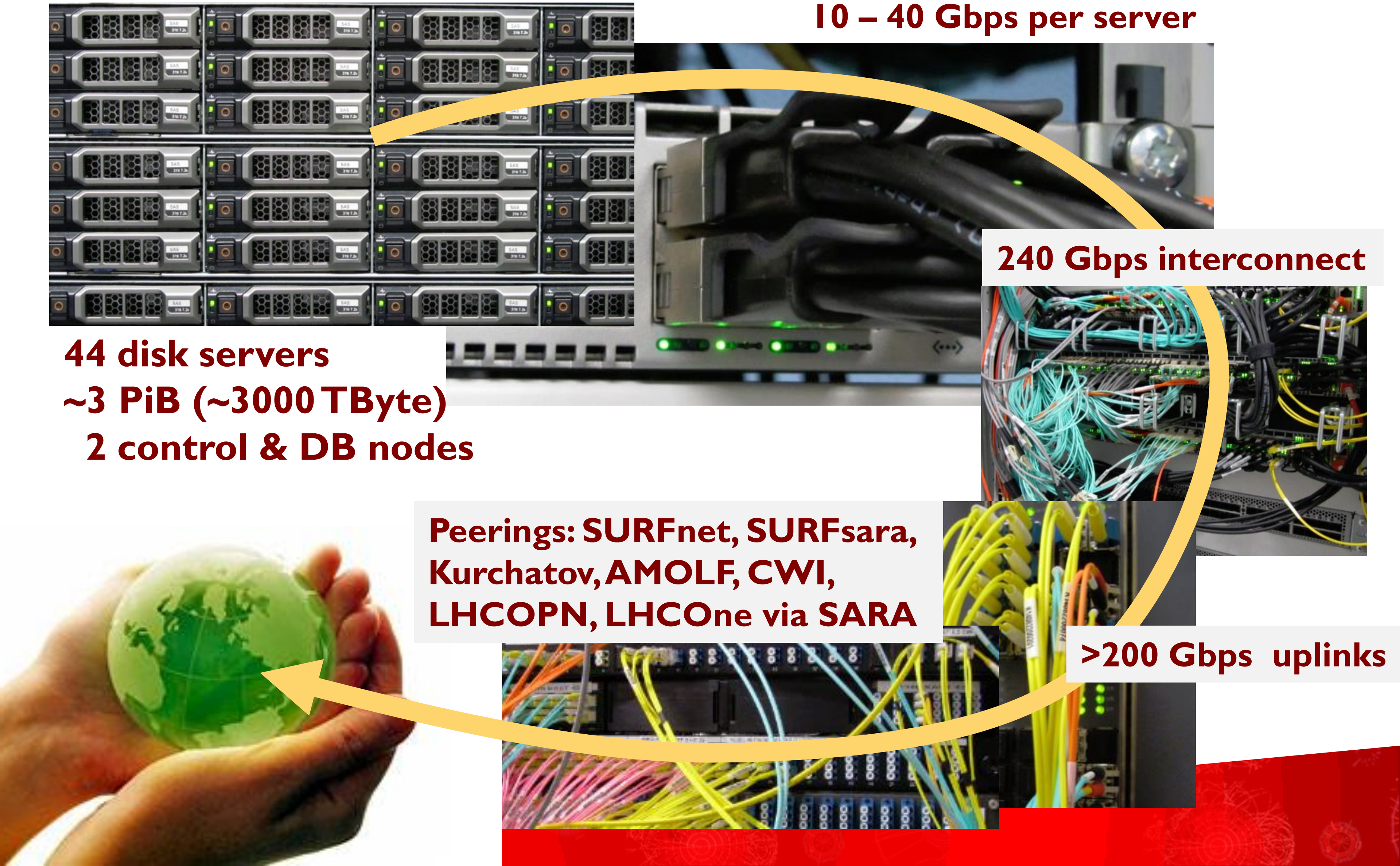
Nik

Am

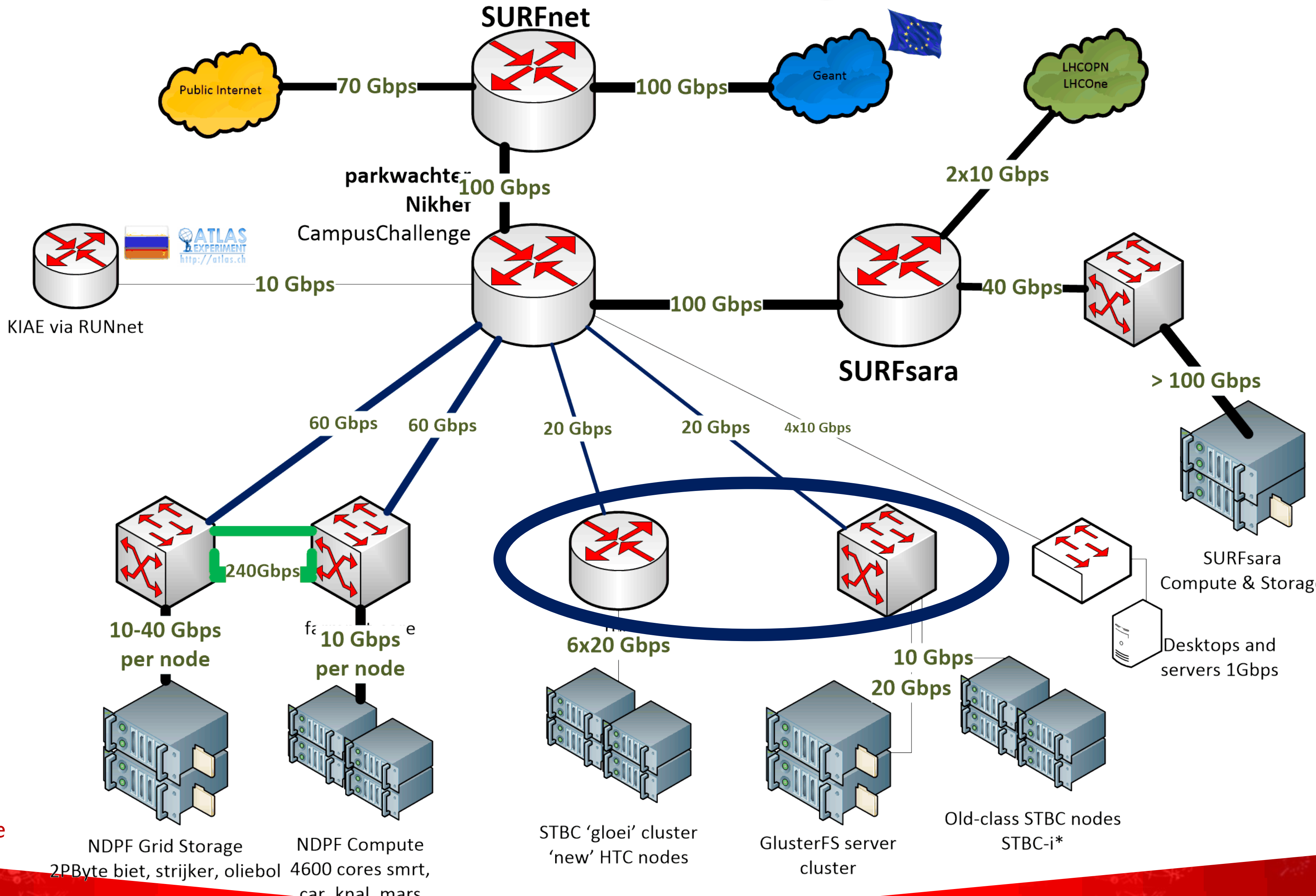
PDI

N

The Flow of Data at Nikhef

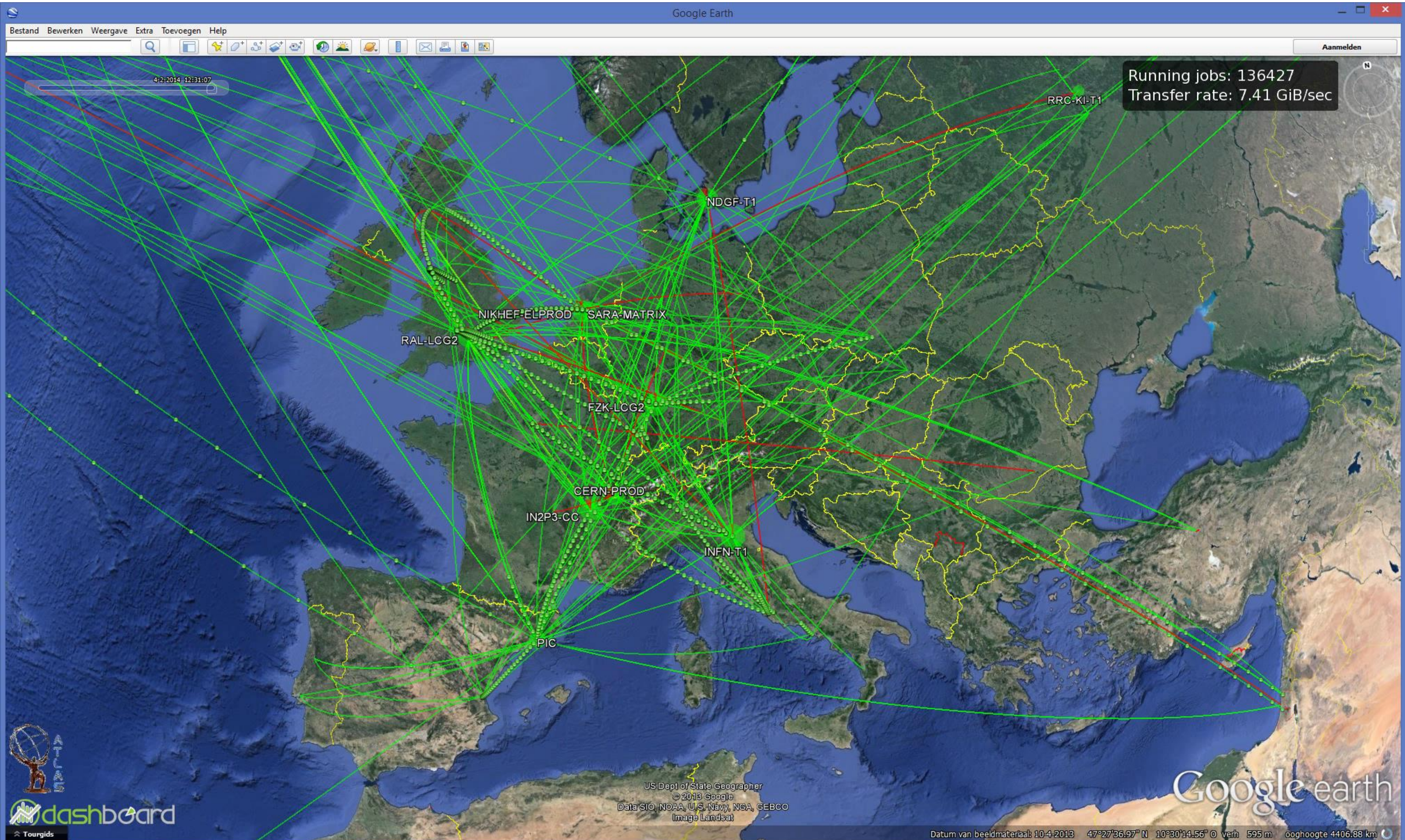


Nikhef Data Processing network



David Groep
Nikhef
Amsterdam
PDP programme

Data Flows in the LHC Computing Grid



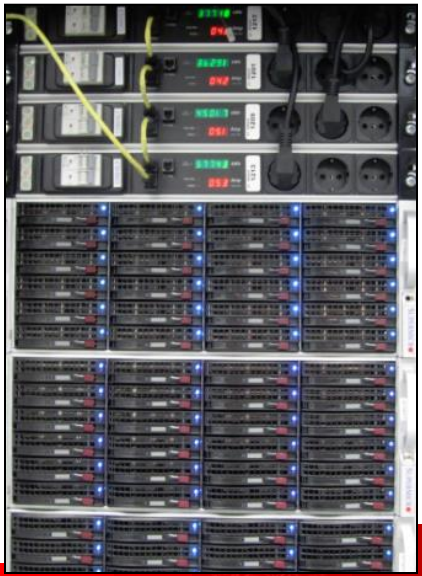
But that's only a small subset

LCG "FTS" visualisation sees but part of the data

- only shows the centrally managed data transfers
- sees only traffic from Atlas, CMS, and LHCb
- cannot show the quality, nor bandwidth used

But each of our nodes sees all its transfers

- server logging is in itself data
- we collect it all



David Groep
Nikhef
Amsterdam
PDP programme

One day worth of logs … ~12GB/day

```
tbn18.nikhef.nl:/var/log/
 631M     dpm/log.1
 1.1G     dpns/log.1
 639M     srmv2.2/log.1
```

plus 44 disk server nodes @250 Mbyte/day

```
09/13 00:01:23.588 26759,79 dpm_srv_proc_put: calling dpm_selectfs
09/13 00:01:23.588 26759,79 dpm_selectfs: selected pool: BIOMED
09/13 00:01:23.588 26759,79 dpm_selectfs: selected file system: oliebol-02.nikhef.nl:/export/data/biomed
09/13 00:01:23.588 26759,79 dpm_selectfs: oliebol-02.nikhef.nl:/export/data/biomed reqsize=0, elemp->free=39997608l749, poolp->free=39997608l749
09/13 00:01:23.645 26759,79 dpm_srv_proc_put: calling Cns_creatx
09/13 00:01:23.712 26759,19 dpm_srv_getspacetoken: DP092 - getspacetoken request by /DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=ddmadmin/CN=53149
09/13 00:01:23.712 26759,19 dpm_srv_getspacetoken: DP098 - getspacetoken ATLASDATADISK
09/13 00:01:23.713 26759,19 dpm_srv_getspacetoken: returns 0, status=DPM_SUCCESS
09/13 00:01:23.753 26759,78 dpm_srv_proc_get: TURL info: gsiftp oliebol-09.nikhef.nl oliebol-09.nikhef.nl:/export/data/atlasprd/atlas/2015-09-12/
09/13 00:01:23.755 26759,78 dpm_srv_proc_get: returns 0, status=DPM_SUCCESS
09/13 00:01:23.761 26759,79 dpm_srv_proc_put: TURL info: gsiftp oliebol-02.nikhef.nl oliebol-02.nikhef.nl:/export/data/biomed/biomed/2015-09-13/j
09/13 00:01:23.763 26759,79 dpm_srv_proc_put: returns 0, status=DPM_SUCCESS
09/13 00:01:23.881 26759,18 dpm_updfreespace: oliebol-02.nikhef.nl:/export/data/biomed incr=0, elemp->free=39997608l749, poolp->free=39997608l749
09/13 00:01:23.881 26759,18 dpm srv putdone: returns 0, status=DPM SUCCESS
```

David Groep
Nikhef
Amsterdam
PDP programme

And then our storage manager is still ‘decent’ …

CASTOR Logs

- ~2GB/day from the node I showed (highest volume)
- ~30GB/day collected overall
- ~200 source nodes
- ~70,000,000 log events/day

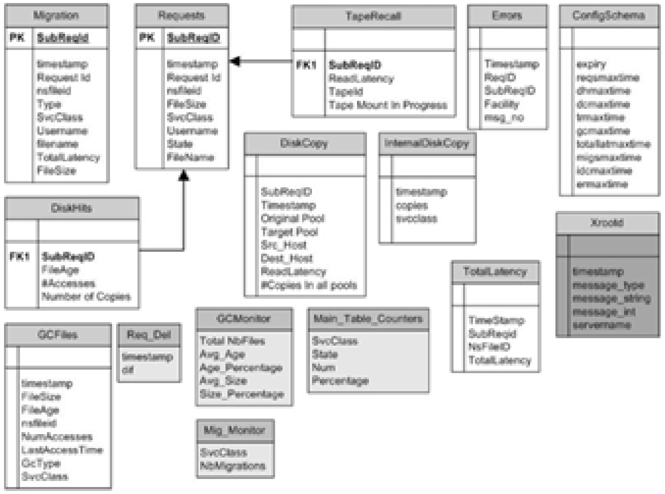
Science & Technology Facilities Council

David Groep
Nikhef
Amsterdam
PDP programme

Nikhef

The First Solution - DLF

- DLF = 'Distributed Logging Facility'
- CERN-developed monitoring system for CASTOR
- Store all the log information in a big Oracle DB



Source: *CASTOR end-to-end monitoring*, by T Rekatsinas et al, URL: http://iopscience.iop.org/1742-6596/219/4/042052/pdf/1742-6596_219_4_042052.pdf

Science & Technology Facilities Council

David Groep
Nikhef
Amsterdam
PDP programme

Running DLF

- Scalability was a killer.
  - By 2013, simple queries were taking >1 hour.
  - Fundamental architecture couldn't cope.

Science & Technology Facilities Council

David Groep
Nikhef
Amsterdam
PDP programme

Big Data Analytics for log analysis

- Analysis of log data is typical 'big data' problem
  - CERN tried Hadoop ('map-reduce')
  - RAL went with … ELK*

- For logs specifically, it's mostly efficient search
  - ElasticSearch (www.elastic.co)
  - LogStash (collect and parse logs, import to ES)
  - Kibana – analysis based on Apache Lucene + graphing

- Integrated into a single 'stack': ELK

David Groep
Nikhef
Amsterdam
PDP programme

*Appleyard et al., http://pos.sissa.it/archive/conferences/239/027/ISGC2015_027.pdf

LogStash

Data arrives in format A…
…process B occurs…
…data out in format C

e.g. convert syslog into json

plugin documentation

inputs

- collectd
- drupal_dblog
- elasticsearch
- eventlog
- exec
- file
- ganglia
- gelf
- gemfire
- generator
- graphite
- heroku
- imap
- invalid_input
- irc
- jmx
- log4j
- lumberjack
- pipe
- puppet_facter
- rabbitmq
- rackspace
- redis
- relp
- s3
- snmptrap
- sqlite

codecs

- cloudtrail
- collectd
- compress_spooler
- dots
- edn
- edn_lines
- fluent
- graphite
- json
- json_lines
- json_spooler
- line
- msgpack
- multiline
- netflow
- noop
- oldlogstashjson
- plain
- rubydebug
- spool

filters

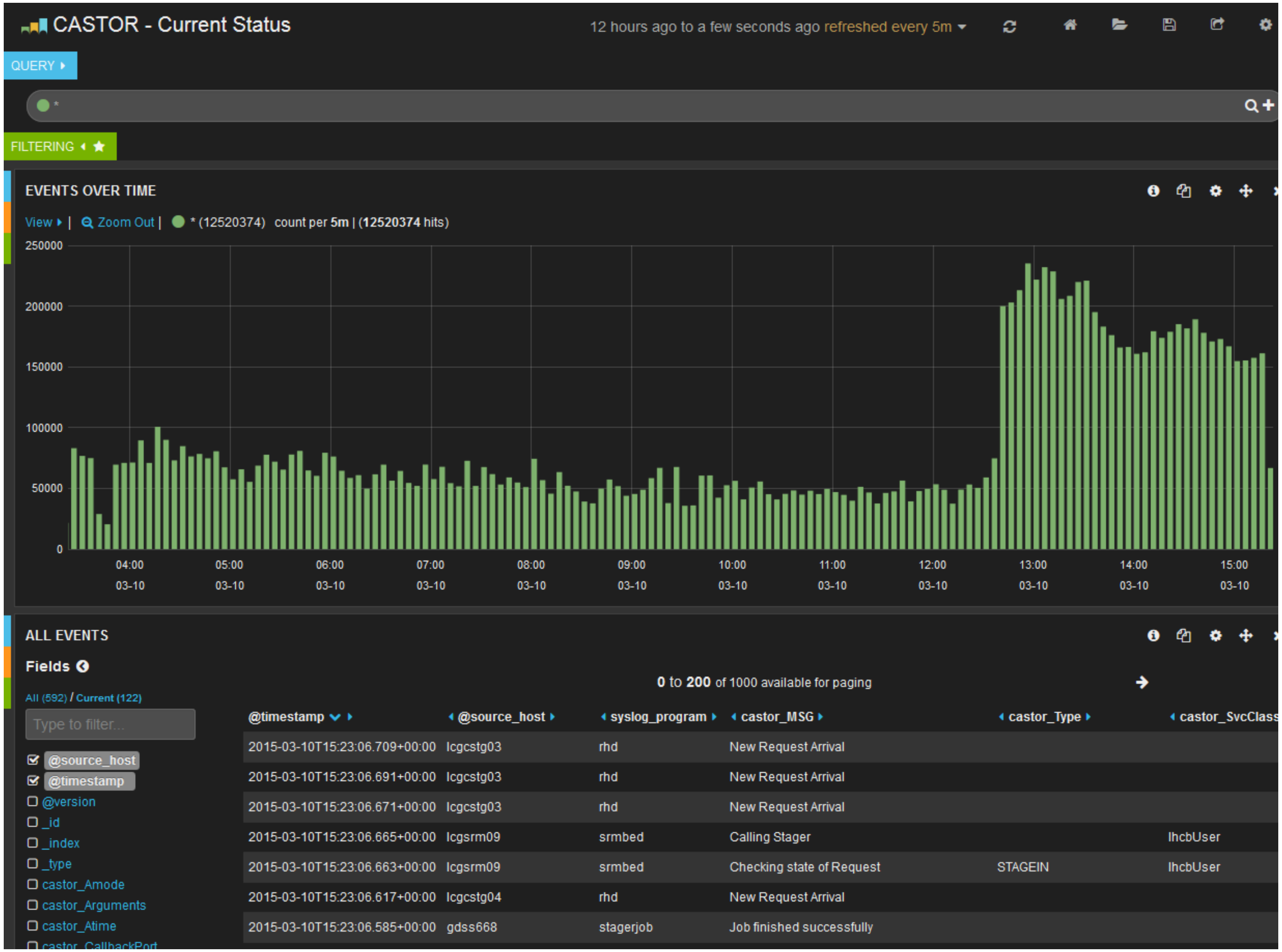
- advisor
- alter
- anonymize
- checksum
- cidr
- cipher
- clone
- collate
- csv
- date
- dns
- drop
- elapsed
- elasticsearch
- environment
- extractnumbers
- fingerprint
- gelfify
- geoip
- grep
- grok
- grokdiscovery
- i18n
- json
- json_encode
- kv
- metaevent

outputs

- boundary
- circonus
- cloudwatch
- csv
- datadog
- datadog_metrics
- elasticsearch
- elasticsearch_http
- elasticsearch_river
- email
- exec
- file
- ganglia
- gelf
- gemfire
- google_bigquery
- google_cloud_storage
- graphite
- graphtastic
- hipchat
- http
- irc
- jira
- juggernaut
- librato
- loggly
- lumberjack

David Groep
Nikhef
Amsterdam
PDP programme

Analyse, for now with Kibana/Lucene



David Groep
Nikhef
Amsterdam
PDP programme

Graphic from Rob Appleyard, RTFC RAL – at ISGC 2015

Challenges ahead!

For now, we have

- 44+ difference data sources, 240 Gbps of traffic
- 150+ different storage partners, 55 countries/regions
- public internet plus the LHCOPN/LHCOne
- 5000+ users, working 24x7

… and 'grep' for a tool … ☹

Phase I:

- setup of a big data analytics cluster (ELK)
- merge diverse data sources into a single system
- define queries and find some global anomalies ☺

David Groep
Nikhef
Amsterdam
PDP programme

Nikhef

Building upon phase I

As an (optional) addition/extension to phase I

- Discuss if ELK indeed the right tool for this: can it cope with the volume? It the ElasticSearch API suitable for defining new visualisations?
- Add additional data sources: are simple transfer logs enough? Does data flow correlate with computing, and can we see that by adding sources and defining (lucene) queries?

And in phase II (>Feb 2016)

- How can global data flows be presented?
- Can one conceive visualisations for the general public? for users? or for both?

David Groep
Nikhef
Amsterdam
PDP programme