CHEP Report

Jeff Templon

NIKHEF pdp

these talks

* Ian's intro talk @ WLCG (link to agenda)
* Daniel van der Ster's user support talk @ WLCG
* Ian's middleware talk
* Les Robertson's plenary
* Andy Hanushevsky's xrootd talk

NIKHEF pdp

Collaboration Workshop

* Much of usual stuff
* New stuff
  * EGI thoughts
  * more details on upcoming run
  * more stuff about actual user support.

NIKHEF pdp

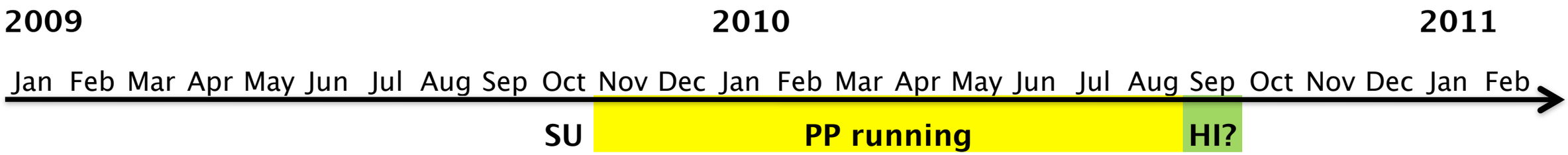
Likely scenario

- ⇨ Injection: end September 2009
- ⇨ Collisions: end October 2009
- ⇨ Long run from ~November 2009 for ~44 weeks
  - + This is equivalent to the full 2009 + 2010 running as planned with 2010 being a nominal year
  - + Short stop (2 weeks) over Christmas/New Year
- ⇨ Energy will be limited to 5 TeV
- ⇨ Heavy Ion run at the end of 2010
  - + No detailed planning yet
- ⇨ 6 month shutdown between 2010/2011 (?) – restart in May ?

- ⇨ Now understand the effective amount of data taking in 2009+2010 will be ~6.1 x 10^6 seconds (cf 2 x 10^7 anticipated in original planning)

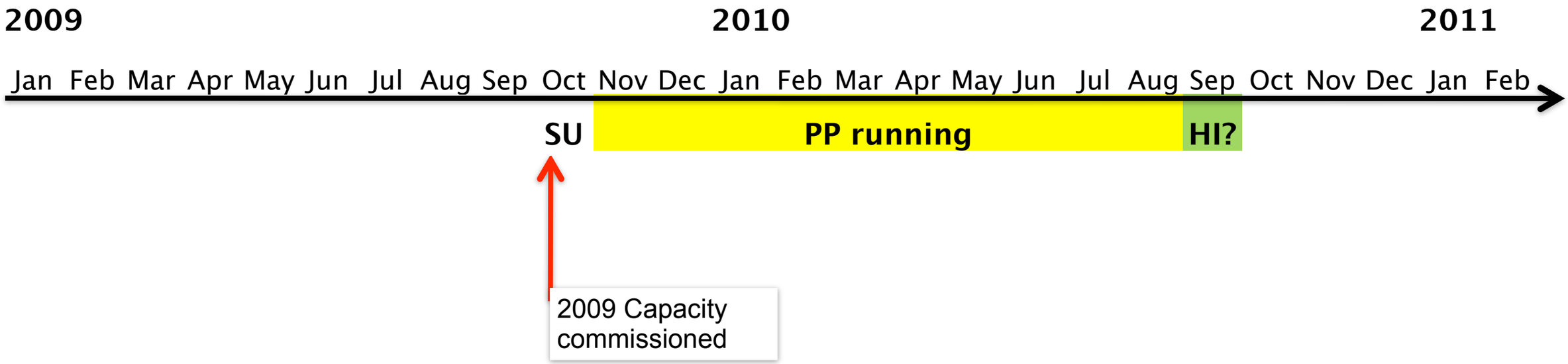
Ian.Bird@cern.ch

WLCG timeline 2009-2010



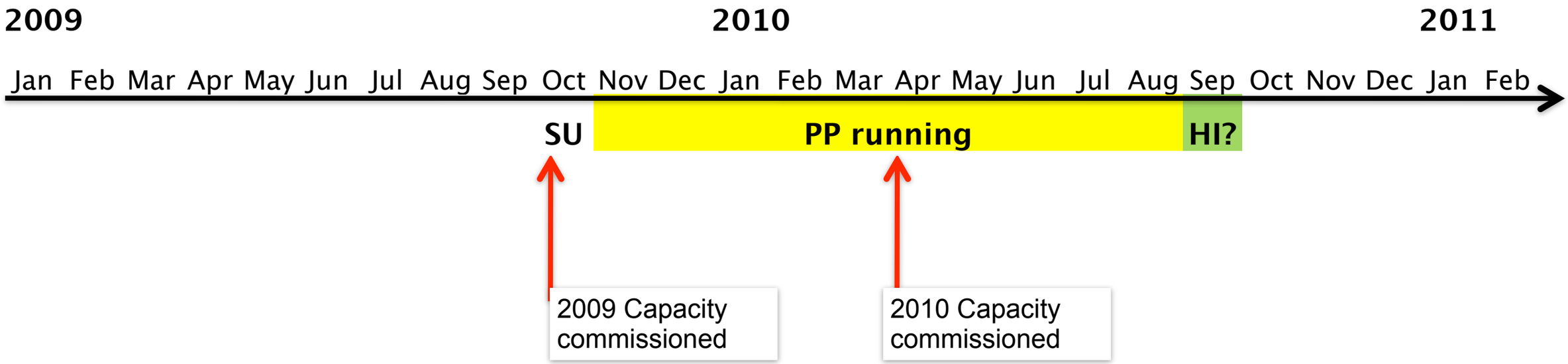
Ian.Bird@cern.ch

WLCG timeline 2009-2010



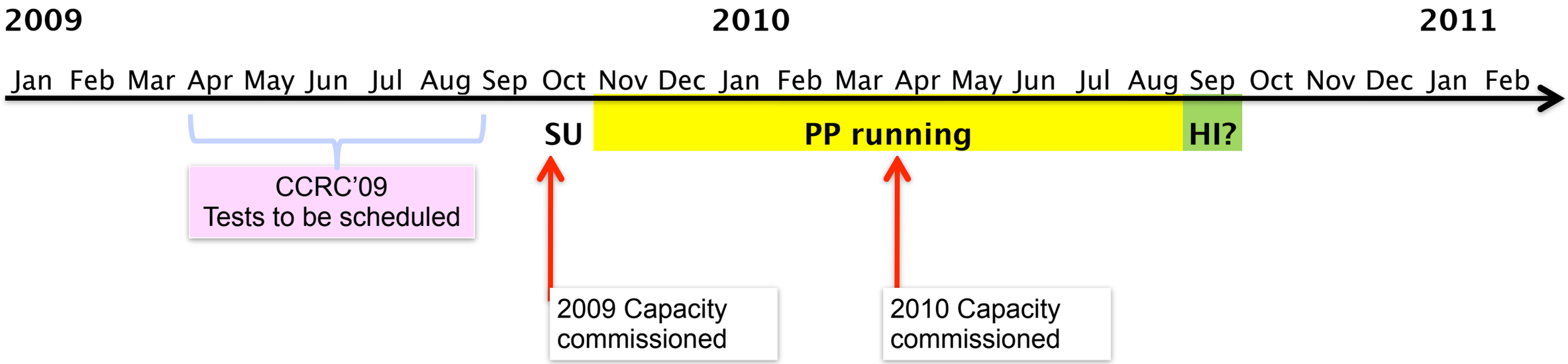
Ian.Bird@cern.ch

WLCG timeline 2009-2010



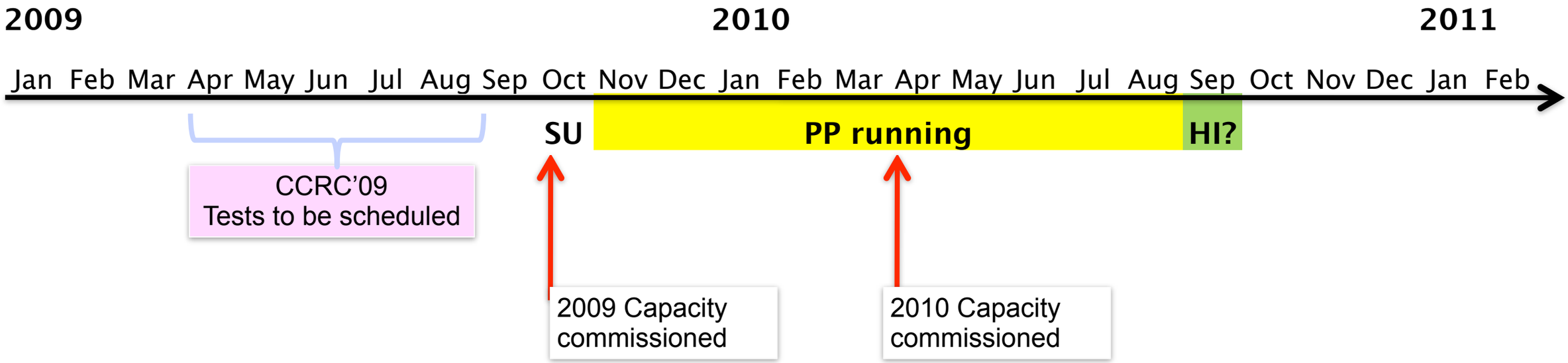
Ian.Bird@cern.ch

WLCG timeline 2009-2010



Ian.Bird@cern.ch

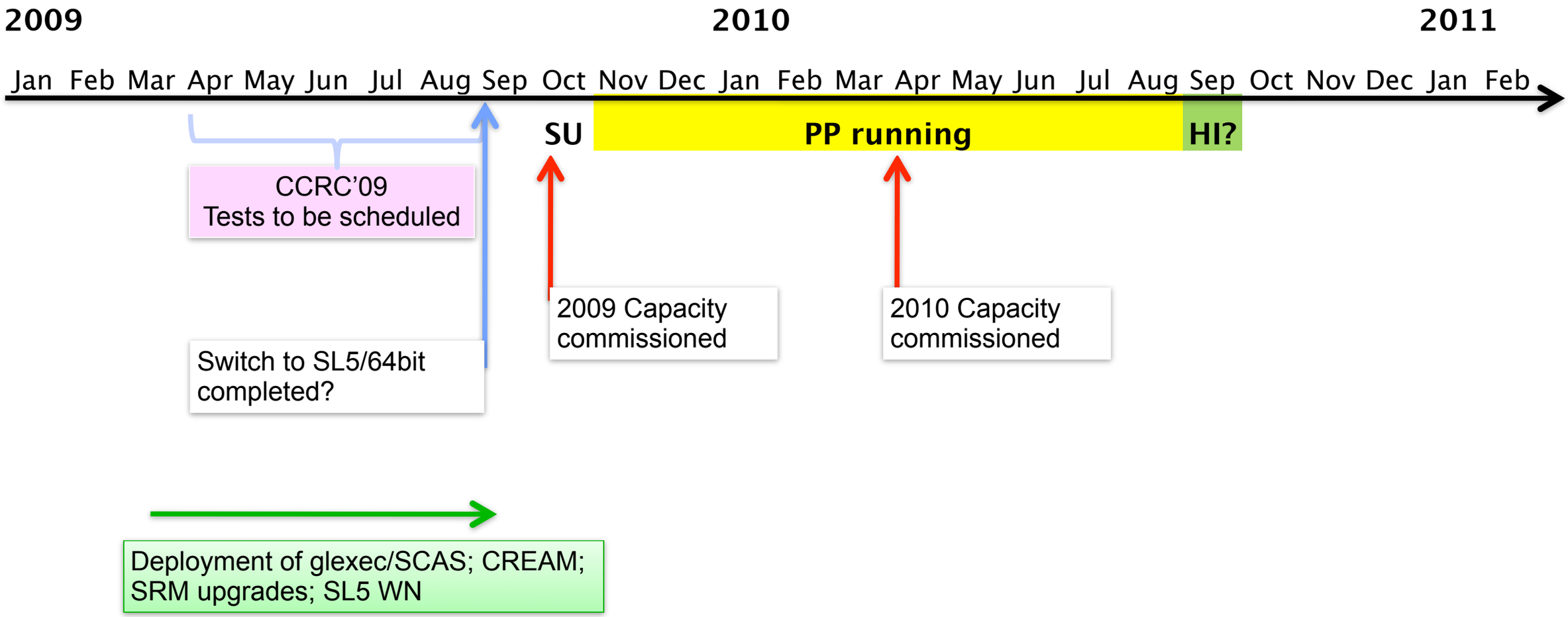
WLCG timeline 2009-2010



Deployment of glexec/SCAS; CREAM; SRM upgrades; SL5 WN

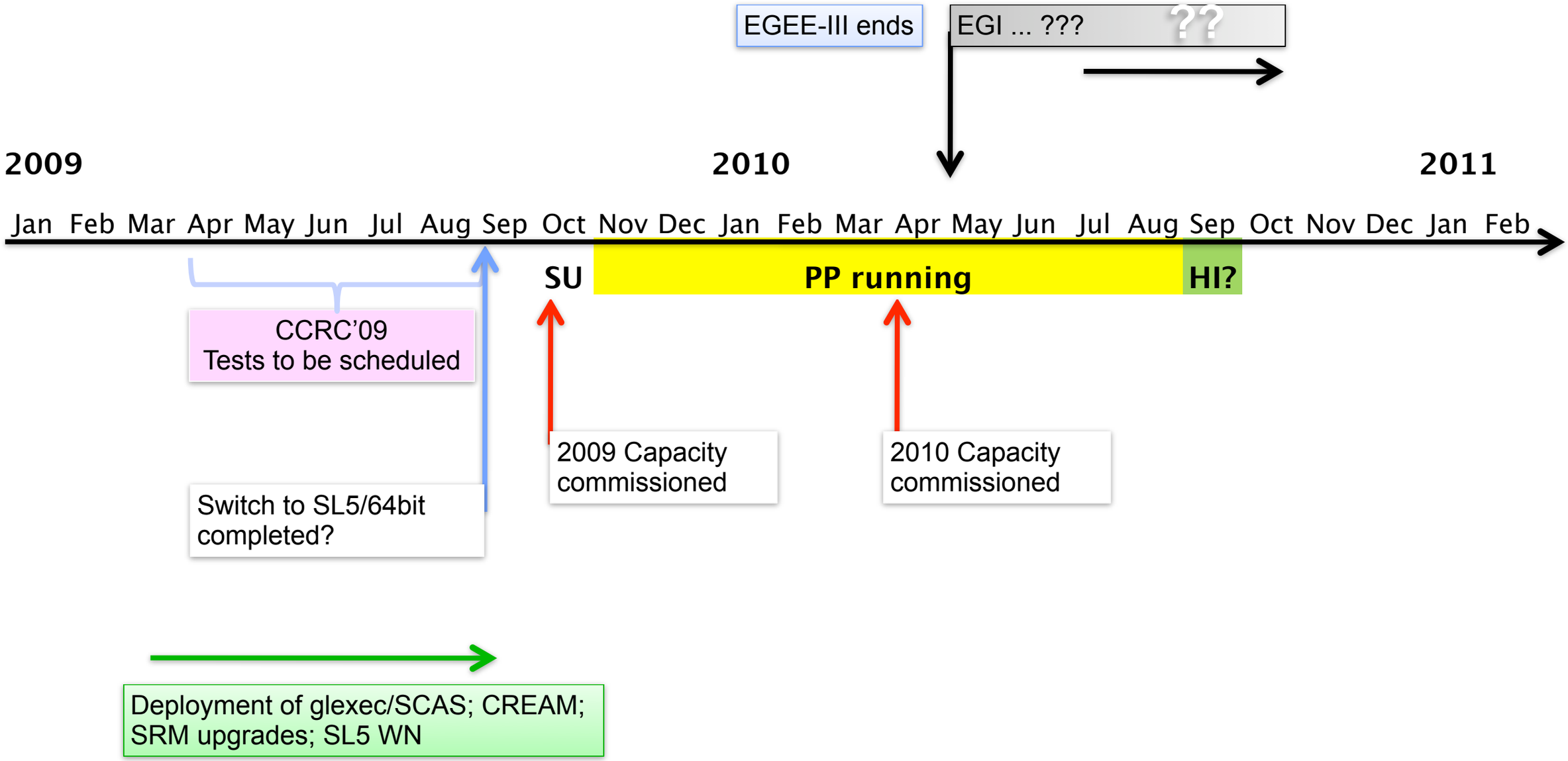
Ian.Bird@cern.ch

WLCG timeline 2009-2010



Ian.Bird@cern.ch

WLCG timeline 2009-2010



Ian.Bird@cern.ch

EGI Workshop in Catania

- Aggressive timescale to take over from EGEE on 1/5/10
  - Incomplete consensus on the set of international tasks (coordination and general services)
- Transition requires milestones to be satisfied:
  - Establish EGI.org (Amsterdam was agreed). Appoint director and staff.
    - Not clear how these appointments can be done
  - Prepare EC proposals for calls closing in Autumn (Nov 09)
    - Need coordinators for these proposals – who?
- For EGEE to proceed with transition plans; need to know which NGIs will be present; which tools will remain etc. Not clear.
- CERN role not clear – does it have a vote?
- Could EGEE-III be extended to help transition?
  - But no more money

Ian.Bird@cern.ch

EGEE Services needed by WLCG (Plan B)

- GGUS
  - Relies on connections to local support ticket systems – today in ROCs and sites
    - → Tier1 and Tier2 sites?
  - COD, TPM
- Operations and Service coordination
  - CERN + EGEE ROCs
- ROCs:
  - Support effort (TPM, COD) → moves to Tier 1s?
- EIS team – CERN (largely LCG funded)
- ENOC
  - Coordination of OPN operations- currently by IN2P3
- Deployment support:
  - m/w deployment/testing/rollout/support
  - Pre-production testing – effort and resources
- Operational Security coordination
- Policy development

- Accounting:

  APEL – infrastructure/DB and service

  NB Italy uses DGAS and publishes into APEL; OSG + ARC publish into APEL

  Portal – CESGA
- GOCDB: configuration DB

  Important for all configurations and definitions of sites and services
- CIC Portal:

  Contact information, VO-ID cards, broadcast tool, Automated reporting,
- Availability/Reliability:

  SAM framework (and migration to Nagios); SAM tests

  Gridview/Algorithms etc:

  GridMap:

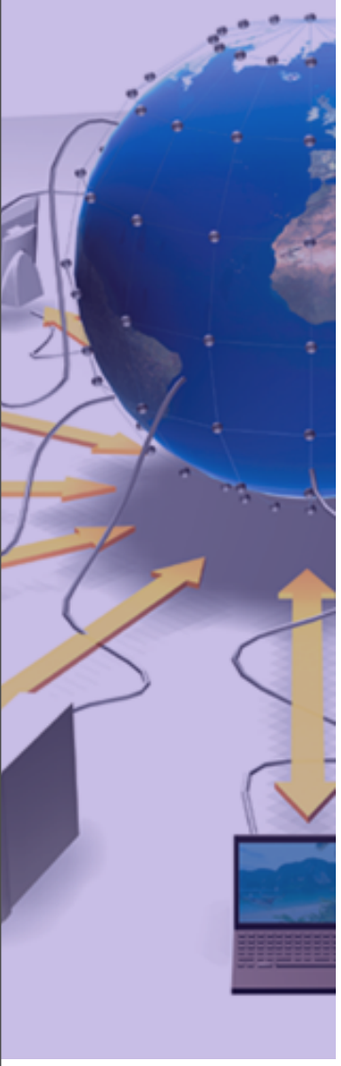
  MSG
- Dashboards

  Service, framework and common services

  Experiment-specifics
- Middleware ...

Ian.Bird@cern.ch

Common Issues



1. Usual DA issues:
   - Why did my job fail? My job ran yesterday but not today?
2. User support is not just DA support
   - The user workflow is (a) look for input data, (b) run the jobs, (c) retrieve the output data
     - Need to support more than Ganga/pathena; (especially data management tools).
3. Users aren't aware of the very nice monitoring:
   - Many users find it more convenient to ask why their job failed, rather than check what the monitoring is showing
4. Users don't (and might never) know the policies:
   - i.e. where they can run, what inputs they can read, where they can store outputs, which storage locations are temporary/permanent, …
   - Policies are dynamic and inconsistently implemented

- 3 & 4 above imply that the end-user tools need to
  - fully enforce the policies, and
  - be fully integrated with the monitoring, especially by being aware of site downtimes

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

owning the middleware

* confidence in "middleware" very low
* observed that most used mw is either CERN or predates EGEE (eg LCMAPS)
* move to independent mw consortia seems to decrease confidence
* not clear how one does it ...

What works?

- Single sign-on – everyone has a certificate, we have a world-wide network of trust
  - VO membership management (VOMS), also tied to trust networks
- Data transfer – gridftp, FTS, + experiment layers;
  - Demonstrate full end-end bandwidths well in excess of what is required, sustained for extended periods
- Simple catalogues – LFC
  - Central model – sometimes with distributed read-only copies (ATLAS has a distributed model)
- Observation: The network – probably the most reliable service – fears about needing remote services in case of network failure probably add to complexity
  - i.e. Using reliable central services may be more reliable than distributed services

What else works

- Databases – as long as the layer around them is not too thick
  - NB Oracle streams works – but do we see limits in performance?
- Batch systems and the CE/gateway
  - After 5 years the lcg-CE is quite robust and (is made to) scales to today's needs ... But must be replaced (scaling, maintenance, architecture, ...). Essentially a reimplentation of the Globus gateway with add-ons
- The information systems – BDII – again a reimplentation of Globus with detailed analysis of bottlenecks etc.
  - GLUE – is a full repository of experience/knowledge of 5 years of grid work – now accepted as an OGF standard
- Monitoring, accounting
  - Today provides a reasonable view of the infrastructure
- Robust messaging systems – now finally coming as a general service (used by monitoring ... Many other applications)
  - Not HEP code!

What about…

- Workload management?
  - Grand ideas of matchmaking in complex environments, finding data, optimising network transfer etc
  - Was it ever needed?
  - Now pilot jobs remove the need for most (all?) of this
  - Even today the workload management systems are not fully reliable despite huge efforts

- Data Management
  - Is complex (and has several complex implementations)
  - SRM suffered from wild requirements creep, and lack of agreement on behaviours/semantics/etc.

And ...

- Disappointment of existing m/w robustness and usability
  - Consistent logging, error messages, facilities for service management, etc....
- Providers have never been able to fully test own services – rely on certification team (seen as bottleneck)
  - Plus problems of complexity/interdependencies have taken a long time to address
- What if WLCG is forced to have its own m/w distribution – or recommended components?
  - Can we rely on a gLite consortium, “EGI” middleware development, etc?
  - How can we manage the risk that the developments diverge from what we (WLCG) need?

Other lessons

- Generic services providing complex functionality for several user communities are not easy
- Performance, scalability, and reliability of basic services are most important (and least worked on?)
- Complex functionality is almost always application specific and should be better managed at the application level
- Too many requirements and pressure to deliver NOW;
  - But lack of prototyping
  - Wrong thing produced, or too complex, or requirements had changed
- Suffered from lack of overall architecture

themes from CHEP

* Energy
* Virtualization
* Clouds
* xrootd, data, data, data

NIKHEF pdp

As LHC starts, data handling depends on a grid
But is the model of a general science grid sustainable?

- WLCG operates on top of multi-science grids
  - With short-term funding cycles <> incompatible with long-term services
  - Hard to find other sciences outside physics that **depend** on these grids
  - Proposal in Europe for a long-term infrastructure (EGI), but still some way from agreement and approval, and EGEE ends next April
    right in the middle of the first LHC run
  - Open Science Grid with a similar role to EGEE in US has 5-year funding from NSF and DoE through 2010
- Could be problematic – but ..
  - Tier-1 sites are still at the heart of EGEE and OSG operations
  - HEP institutes, collaborators and experiments are to a large extent responsible for the middleware
- So WLCG and LHC funding agencies can and **surely will** take on the necessary operational responsibility if EGEE and/or OSG close down

1998 – Can you remember that far back?

- Clinton was President and the Lewinsky scandal broke in January
- DEC was bought by COMPAQ, which has since been absorbed by HP
- Amazon had been selling books online for only three years
- The Google name was not registered until September

- Intel introduced the Pentium II on 250 nm technology, with 7.5M transistors, and up to 450 MHz clock (2008 – Core 2 – 45 nm, 410M transistors, 3.2GHz)
- Top of the TOP 500 list was the ASCI Red at Sandia with 9,152 processors delivering 1.3TFLOPS (2008 – a 129,600 core system at Los Alamos delivering 1,105 TFLOPS)
- The European Research Network backbone was upgraded from 34 to 155 Mbps in December
- 64 kbps was an excellent home network connection
- WiFi prototypes were just appearing (IEEE 802.11-1997)
- and GPRS data services on GSM phones were yet to be launched

Energy

- “Data centres consumed 1 per cent of the world’s electricity in 2005. By 2020 the carbon footprint of the computers that run the internet will be larger than that of air travel, a recent study by McKinsey and the Uptime Institute predicted.”

  Times Online – September 2008

- Even if the cost of oil is down at ~$50, if this growth rate really continues – **power-efficient data centres and cheap renewable energy** must be essential components of any infrastructure that is being planned today

☺**The distributed (grid) model enables us to incorporate data centres wherever they may be located, and whoever is running them**

Virtualisation re-born

- A virtual machine implements a set of low-level functions
  - That can support a full-blown operating system such as Linux
  - And can be mapped on to a wide range of hardware
- The real operating system is reduced to being a virtual machine manager
- While the application is packaged together with the full function operating system and everything else it needs

→ The application decides which software configurations to support, when to upgrade

→The computer centre is free to share processors between different applications,
and to install new hardware on its own timescale

→No longer have to take global decision on versions of OS + middleware + ..+.. +.. freedom from the **dependency gridlock**

No free lunches

- While the applications are now independent of each other they still need the functionality of Linux, software, middleware, and all the other external packages
  → integration, certification, test beds, ….
- Reduced scope for hackers? No more critical security upgrades??
  Security no longer the responsibility of the computer centre???
- Can you really remove the gridlock and retain inter-operability?

  … and maintainability?
  The reason CERN standardised on Linux+Intel (exactly 10 years ago) was to avoid the cost of supporting 7 or 8 flavours of Unix

CernVM Project

- Portable Analysis Environment using Virtualization Technology (WP9)

- Project goal:
  - Provide a complete, portable and easy to configure user environment for developing and running LHC data analysis locally and on the Grid independent of physical software and hardware platform (Linux, Windows, MacOS)
    - Decouple application lifecycle from evolution of system infrastructure
    - Reduce effort to install, maintain and keep up to date the experiment software
    - Lower the cost of software development by reducing the number of compiler-platform combinations

- Approved in 2007 (2+2 years) as R&D activity, started January 2008

Virtualisation

- Maintenance of the application environment
- Efficient utilisation of farms of multi-core processors
- Portability of applications between operational environments

☺If this technology had been practical ten years ago we would surely have built the grid on it

☺The CernVM approach looks like a very good direction –
start with end-user analysis, lightweight, platform-independent, maintainable model

Cloud v. Grid

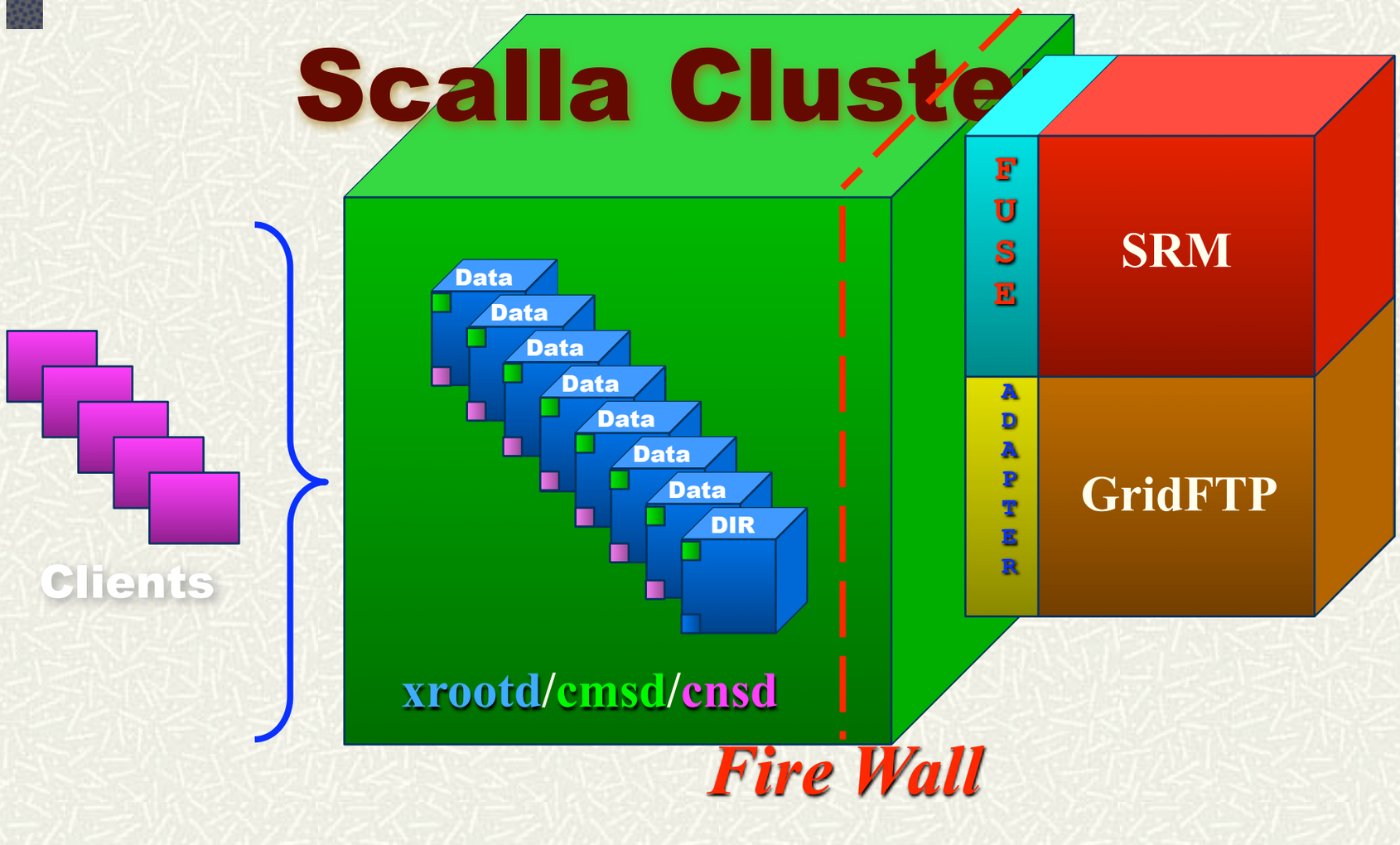
- Clouds aim at efficient sharing of the hardware
  - low-level execution environment, Isolation between users
  - Operated as a homogeneous, single-management domain
  - Straight-forward i/o and storage
  - Expose only a high-level view of the environment – scheduling, data placement, performance issues are hidden from the application and the user
- Grids aim at collaboration
  - Add your resources to the community, but retain management control
  - Expose topology – location of storage, availability of resources
  - Choice of tools to hide the complexity from the user, and the application can write its own tools
- **Both need complex middleware to function**
  - Grids had a problem in trying to provide a universal high-functionality environment (OS, data management, ….), with intersecting collaborations and a naturally competitive environment
  - Clouds have an advantage in offering a simpler base environment, leaving much of the functionality to the application – where universal solutions are not necessary – and what they do have to provide can be decided within a single management hierarchy
- As the names suggest –

  the grids are transparent and the clouds are opaque

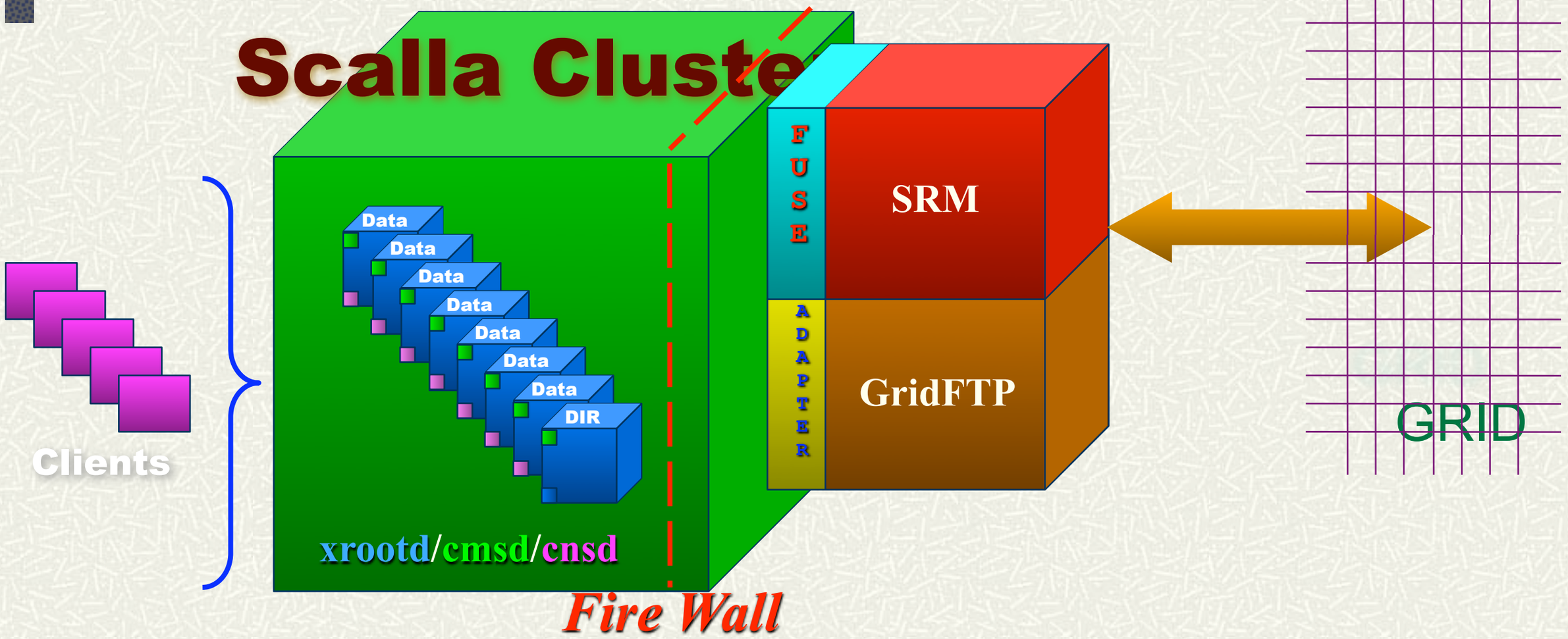
.. and Mobility

- ADSL at 20 Mbps, WiFi/WiMAX/3G
  - We are close to having good bandwidth data connections almost everywhere we go
  - And we already have a powerful high capacity computer in the backpack
- **This is where end-user analysis is going to be done**
- The physicist's notebook must be integrated with the experiment environment, the physics data, and the grid resources
- Without burdening the notebook or its user
- The grid environment is too complex to be extended to the notebook

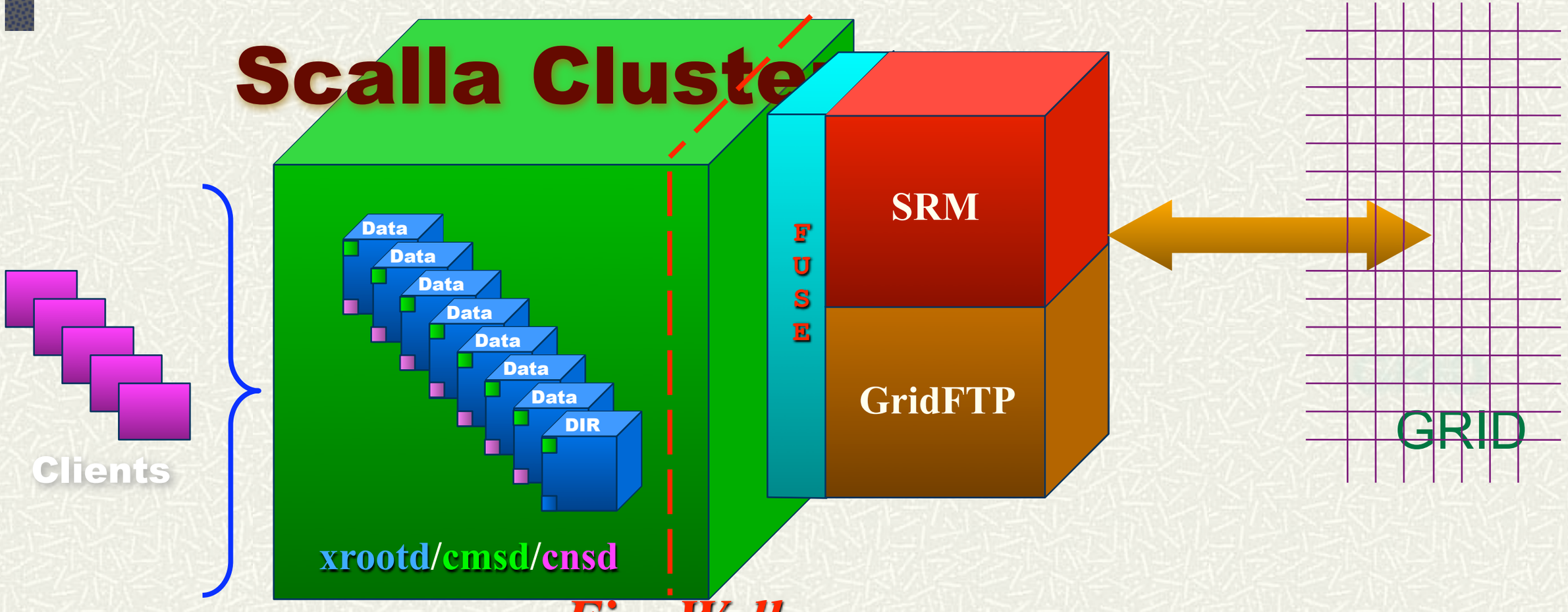
The Scalla/xrootd SE



The Scalla/xrootd SE



The Scalla/xrootd SE



Fire Wall

But wait!

Can't we replace the source adapter with the target adapter
Why not use FUSE for the complete suite?

Integration Recap

- GridFTP
  - Using POSIX preload library (source adapter)
- SRM (BeStMan)
  - Cluster access using FUSE (target adapter)
  - srmLS support
    - Using distributed cnsd's + central xrootd processes
  - Static space token support
    - Using the built-in xrootd partition manager

Because Simpler May Be Slower

- Currently, FUSE I/O performance is limited
  - Always enforces a 4k transfer block size
  - Solutions?
    - Wait until corrected in a post 2.6 Linux kernel
    - Use the next SLAC xrootdFS release
      - Improved I/O via smart read-ahead and buffering
    - Use Andreas Peters', CERN xrootdFS
      - Fixes applied to significantly increase transfer speed
    - Just use the Posix Preload Library with GridFTP
      - You will get the best possible performance

other very interesting talks

* Scalla/xrootd WAN globalization (Furano)
* CMS Software performance (Elmer)
* End-to-end monitoring for DM (Lemaitre)
* Entire session 10 on Monday (Fabrics)

NIKHEF pdp

more

* plenary session 2 on tuesday (Data center evolution!!)
* CERNVM (Buncic)
* VMs plus batch systems (Oberst)
* Oliver Keeble's Middleware talk
* Solid State Drive performance (Panitkin)

even more

* many-core apps (Innocente)
* distributed data analysis (Mato)
* monte carlo in amazon EC2 (Sevoir)

NIKHEF pdp