

Die Grid-Infrastruktur und die NAF am DESY

- Aufbau Grid und NAF, aktuelle Ressourcen und Zahlen
- Erfahrungen mit erster Analyse von LHC Daten
- Einige geplante und mögliche Neuerungen

Yves Kemp für DESY Grid und NAF
DPG Bonn
18.3.2010

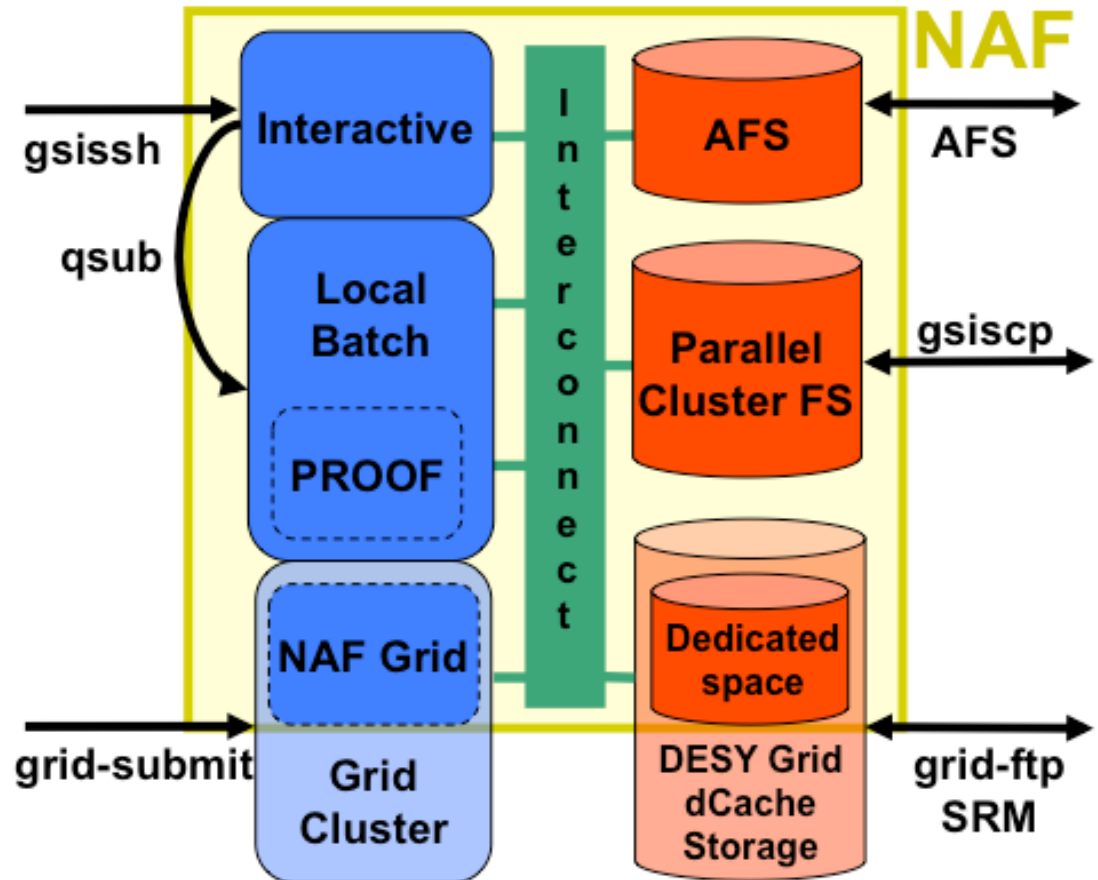
DESY & Grid

- Verteilt auf die beiden DESY Standorte, Produktion seit Ende 2004.
- Bisläng hauptsächlich Experimente an Beschleunigern
 - HERA: H1, ZEUS, HERMES, (HERA-B)
 - LHC: ATLAS, CMS und LHCb
 - Linear Collider: ILC & Calice
- Aber auch neue Communities:
 - BioMed
 - Astroteilchenphysik: IceCube, CTA
 - Photon Science (Petra3, FLASH, DORIS, CFEL, XRAY von ESRF, XFEL.EU, ...)
- Vollständige Grid Infrastruktur: Neben CPU und Storage (eines der größten Tier-2 im LHC) auch alle VO-Services (VOMS, LFC, WMS...)



Und die NAF?

- > Die NAF ergänzt die DESY Grid Ressourcen.
- > Einige Grundideen:
 - Nutzer von deutschen Instituten
 - Für Atlas, CMS LHCb und ILC/Calice
 - Zugang zu allen relevanten Analysedaten
 - Alternative Zugänge zu diesen Daten
 - (Zusätzliche Hardware)
 - Zusätzliche Analysetools
 - Zusätzlicher Support durch Betreiber und Experimentgruppen



Lose Sammlung von Betriebserfahrungen

Lektionen aus dem Atlas HammerCloud

- > ZN: Sehr gute Performance bei HammerCloud: CE und SE gut abgestimmt
- > HH: Performance eher mittelmäßig. 10 Gbit Upgrade der Pools steht noch aus, wird uns viel Performance bringen.

Ein prinzipielles Problem (ein Beispiel):

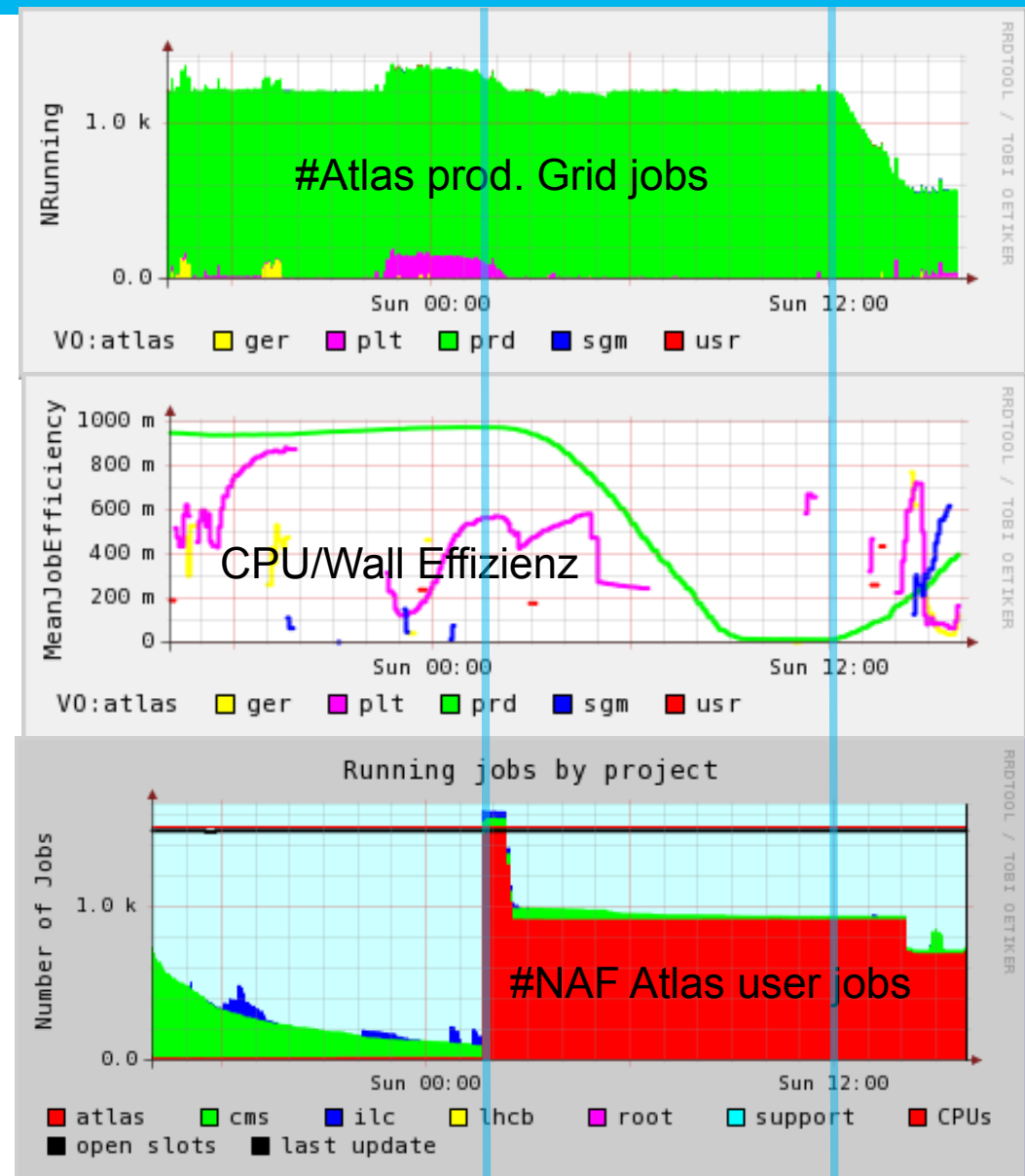
- > Ein grosser CPU Cluster (Grid) und NAF bedient drei dCache Instanzen
 - > Wegen ungleichem Submitverhalten kommt es zu Spitzen in der Auslastung bei einer VO: Kein Problem bei MC-Jobs
 - > Allerdings ein Problem bei Analysejobs: Zu viele Jobs lesen dann vom gleichen Storage
- > Gesehen hauptsächlich bei ATLAS, z.B. den Hammercloud Tests
 - > ~300 Jobs sollten laufen, tatsächlich gelaufen sind ~1000
 - > Atlas SE war überlastet ... aber prinzipiell stabil



Betriebserfahrung (2): Queueing, Variante

Queueing, Variante

- NAF und Grid Cluster lesen gemeinsam vom Atlas HH-SE
- Aus NAF kamen (zu) viele Requests, dcap-Türen blockierten.
- Hat diesmal auch die Grid-Produktion betroffen
- **Wunsch: Kategorisierung Jobs**
 - Datenintensiv
 - Überwiegend CPU
- **Sehr hilfreich!**
- ... aber wie implementieren?



Betriebserfahrung (3): NAF und Lustre

> Gemischte Erfahrung

- Lustre kein einfaches Produkt
- Wir haben keinen kommerziellen Support: Support durch die Community am ehesten wenn man die aktuellste Lustre Versionen fährt
- Jeder Upgrade an den Servern bedeutet kompletten (!) Shutdown aller NAF Rechner

> Nutzungsprofile

- Metadatenperformance schlecht (bekanntes Problem). Bei uns verschärft dadurch, dass User teilweise Releases und Logfiles (Ganga) in Lustre haben. Versuchen wir zur Zeit durch zusätzlichen AFS Platz abzumildern
- ATLAS will automatisches Löschen nach 30 Tagen nicht-Lesen. Klingt einfach, ist aber nicht trivial... und kann vermutlich auch leicht ausgehebelt werden. Wir sind aber dran.

> Mögliche Alternativen:

- Wir sehen keine wirklichen auf kurze Sicht...
- Erstmal bleiben wir bei Lustre (und wir werden auch besser)

Betriebserfahrung (4): Kommunikation mit Usern

> EGEE: Alles einfach:

- Downtime und sonstige Benachrichtigungen an VO Admins und Usern
- Kommunikation GGUS Tickets im Fehlerfall

> Realität etwas anders:

- EGEE-Broadcast zu mächtiges Instrument, manchmal möchte man nur eine Subgruppe erreichen
- Manche User erreicht man nicht per EGEE-Broadcast, andere kommunizieren nicht per GGUS, sondern per privater Email.
- Kommunikation Site → User bei Denial of Service schwierig
- VO-interne Kommunikationswege teilweise orthogonal zu GGUS und EGEE Broadcasts

> NAF: Wir sind modern ☺

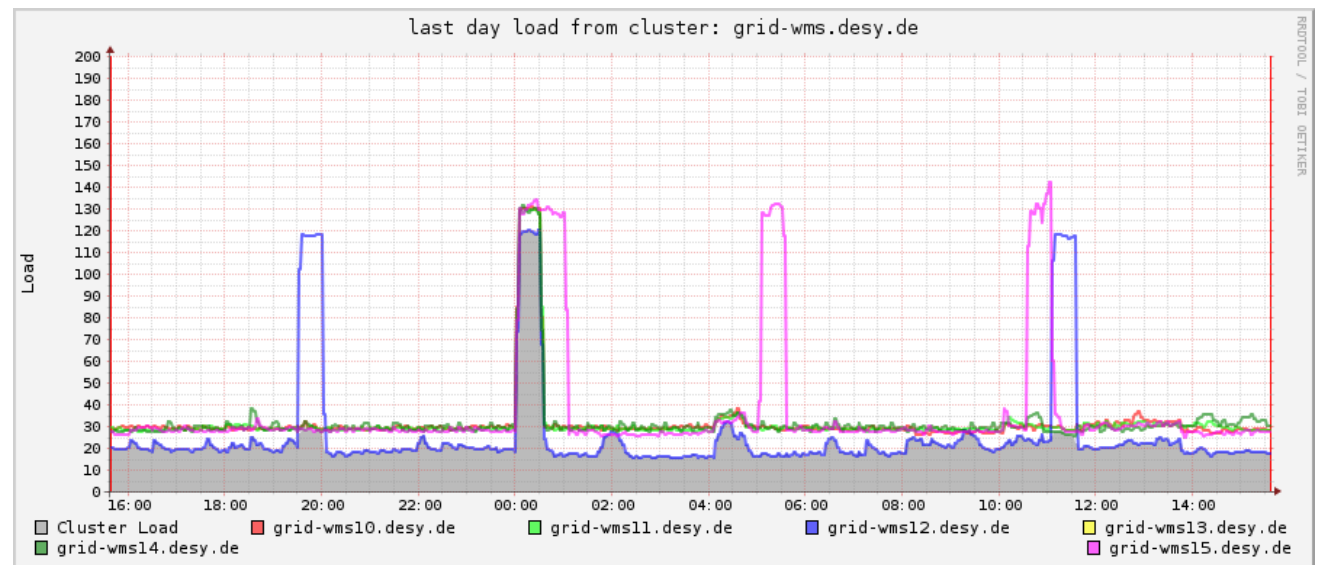


- http://twitter.com/NAF_DESY
- Soll mehr Informationen übermitteln als die traditionellen Wege.
- Erstmal ein Versuch, wir brauchen auch hier vermutlich noch mehrere Iterationen



Betriebserfahrung (5): WMS

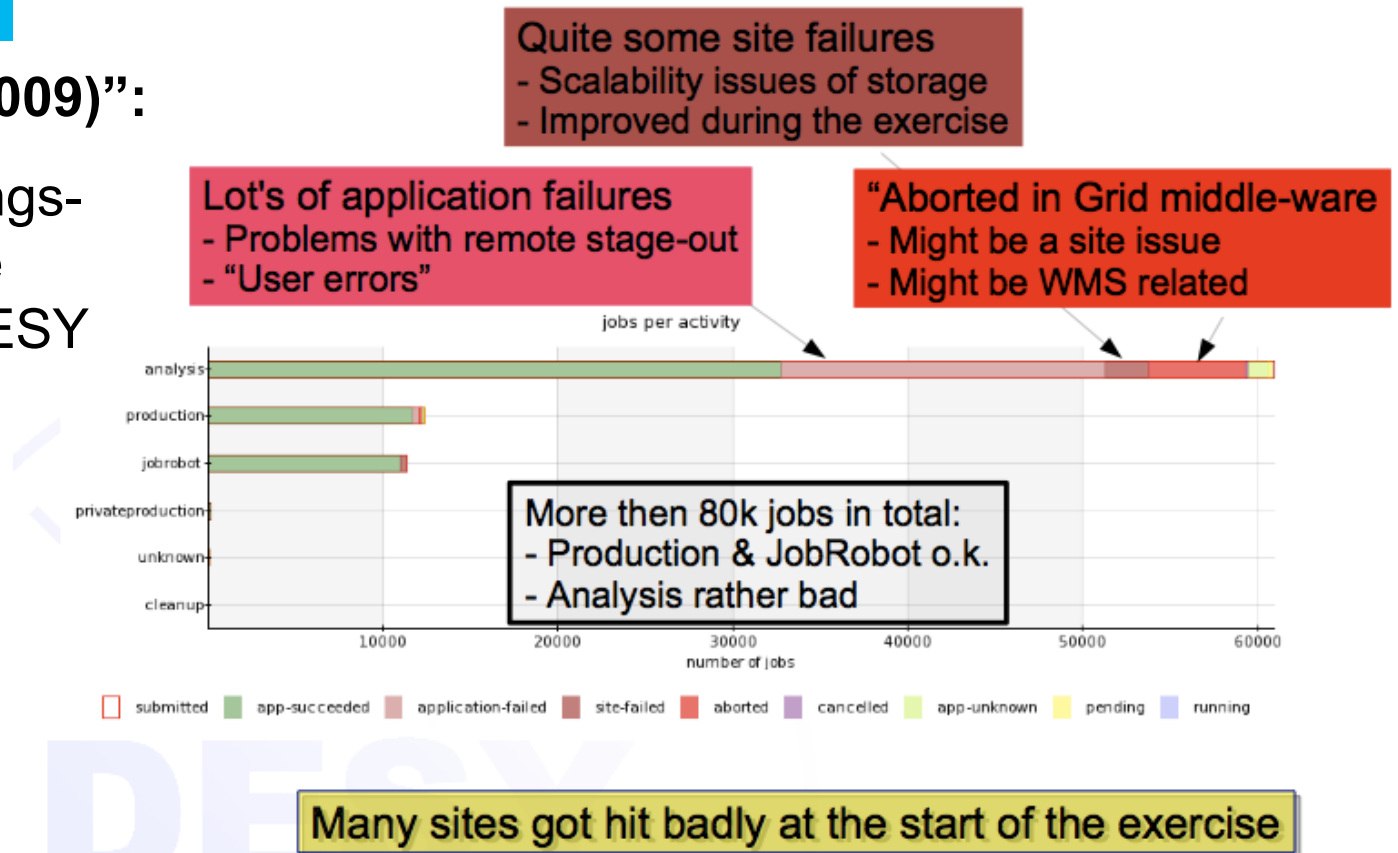
- WMS: Brauchen recht viel Betreuung
 - Stabilität und Administrierbarkeit nur unwesentlich verbessert im letzten Jahr
- Loadbalancing mittels DNS hilft sehr
 - Intelligentes Loadbalancing: Fehlgeschlagene Selbsttests treiben die Load automatisch in die Höhe, so dass kein zeitnaher Eingriff erfolgen muss
 - Unverstandene Probleme bei einer VO mit Job Collections
 - Trotzdem unhandlich.





“October Exercise (2009)”:

- Nach ein paar Anfangsschwierigkeiten gute Performance von DESY



2.36 TeV run:

November, 2009

Christoph Wissing - DESY

8

- Fünf Stunden nach Ende der Datennahme: Alle Daten in DESY-HH und erste Analyse mit Anti-KT Jet-Algo fertig
 - (Sebastian Naumann-Emme und Uni-HH CMS Gruppe)



Neuerungen in der Middleware

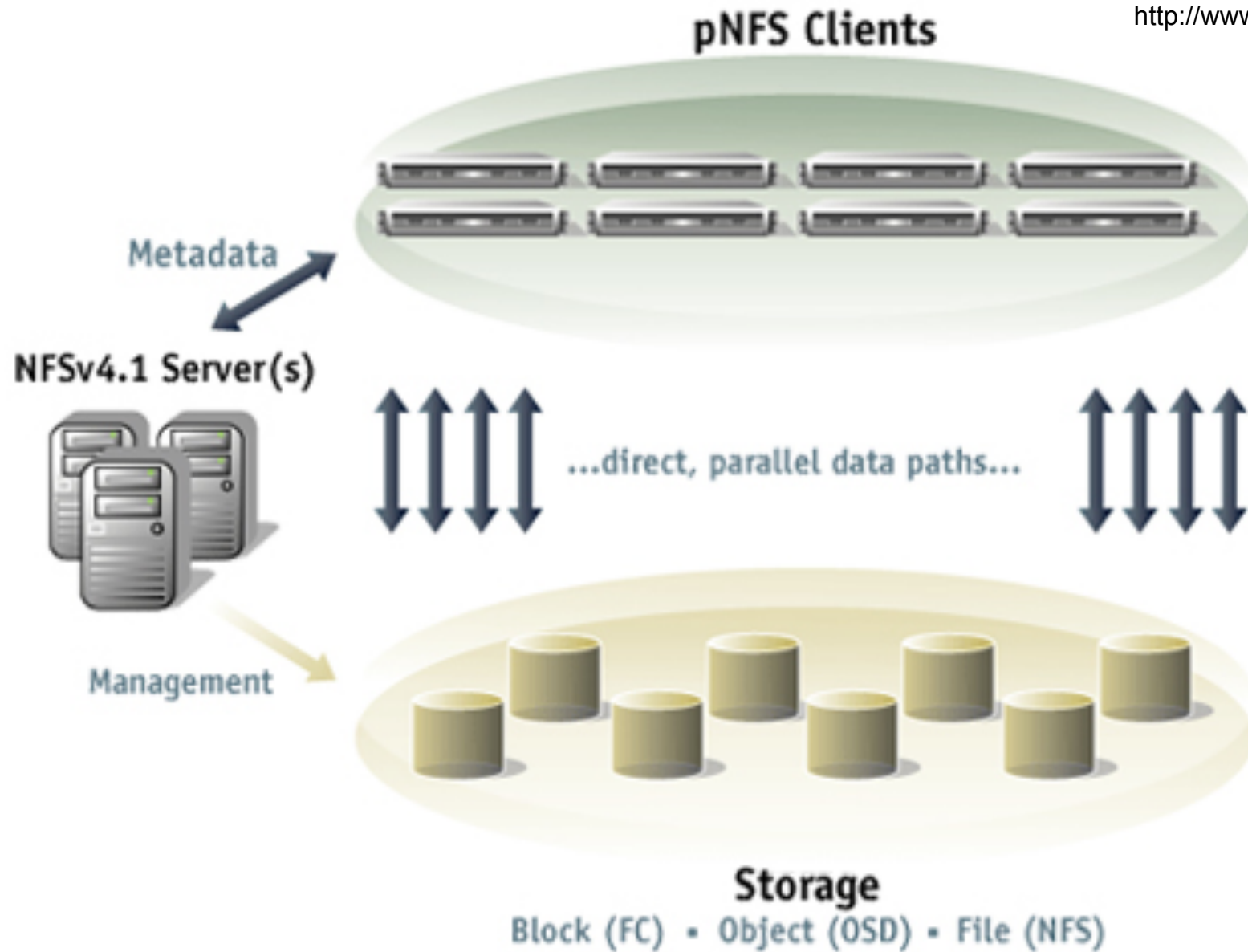
Pilotjobs und glExec

- > Pilotjobs zur Produktion: Kein wirklicher prinzipieller Einwand
 - Pilot ist hier nur ein etwas anders gearterter Wrapper um die Submitskripte
- > Pilotjobs zur Analyse: Hebelt Site-Scheduling aus. (Siehe auch vorherige Folien)
 - Führt zu allen möglichen Problemen.
 - ATLAS hat “nationale” Pilot-Factories eingerichtet, hilft gegen einige Beschwerden.
- > glExec, SCAS, ARGUS,.....:
 - Paradigmenwechsel: Interface zwischen Grid und Site ist nicht mehr das CE, sondern der WN.
 - Wirft viele prinzipielle Fragen auf, von lokalem Fabricmanagement über Accounting bis hin zu Security.
 - Wir hätten uns kurz vor Datennahme andere Diskussionen gewünscht
- > CREAM-CE: In Testing-Phase, wird kommen



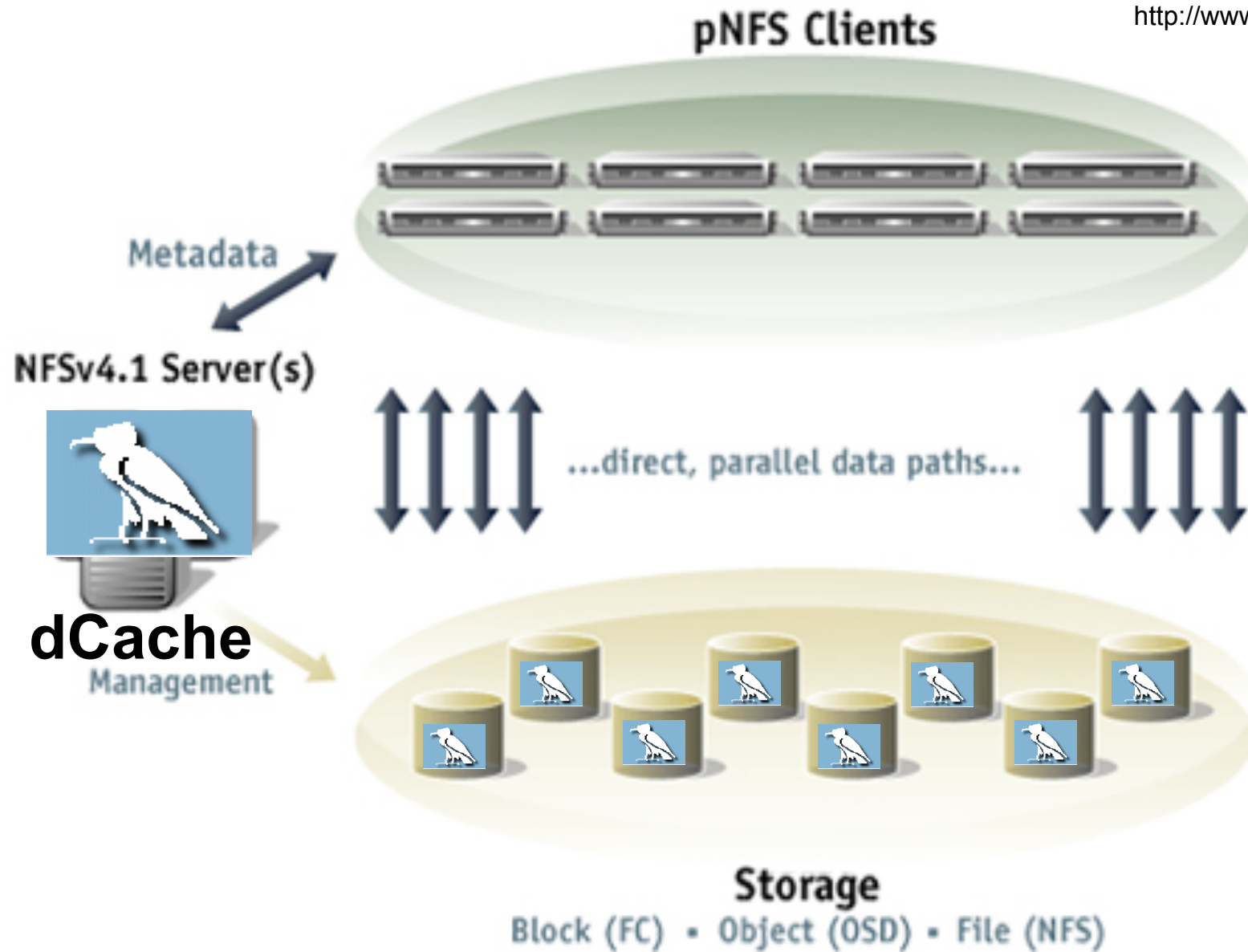
NFS 4.1, pNFS

<http://www.pnfs.com/>

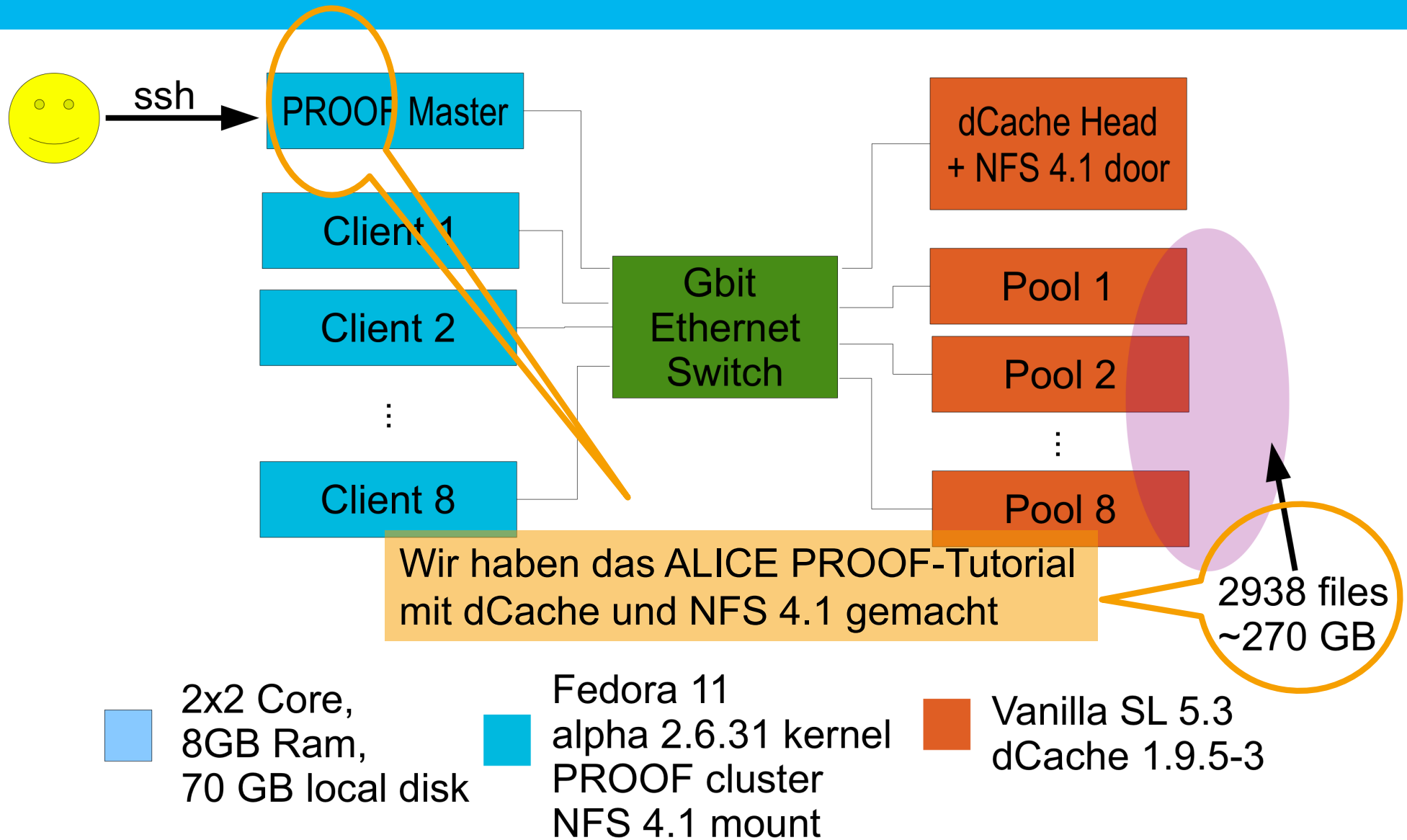


NFS 4.1, pNFS und dCache

<http://www.pnfs.com/>



Erste real-life Versuche im DESY Grid-Lab



gezeigt auf der HEPIX Fall 2009, Demo unter

http://grid.desy.de/gridlab/NFS41/dCacheNFS41_AlicePROOF_HEPIX_movie.swf



Planung von NFS 4.1

- > Sehr interessante Technik: NFS 4.1 passt vom Design her gut zu dCache, und ist ein Industriestandard
- > Warnung: Vermutlich nicht 2010 in Produktion (leider)
- > Wir bereiten das aber vor, verschiedenes muss zusammen kommen:
- > Kernel: Support für pNFS (NFS 4.1): 2.6.36 geplant, Ende 2010
- > Vorher: Selbst gepatchte Kernel. Erster Versuch nur mit Fedora 12, jetzt auch mit SL 5.3 als Basis. Hardware aus dem “Grid-Lab”
- > Weitere Tests / Debugging / Entwicklung auf Client und dCache-Seite
 - In Vorbereitung: Photon Science Anwendung, andere Anwendungen (HammerCloud?)
- > Einsatzgebiete:
 - Vereinfachung des Userzugriffs auf den dCache
 - NAF und Computing für die Photon-Community
 - Ersatz für den altbekannten und /pnfs-Mount (Achtung: Der war NFS v2 und dcap)



Zusammenfassung, Ausblick

Einige (vermutlich doch kleinere) Probleme, generell aber:

- > DESY und NAF bereit für Analyse von LHC Daten
 - Erste Erfahrungen mit Analyse OK bis sehr gut

- > Mit NFS 4.1/pNFS für dCache sehr interessante Aussichten am Horizont

