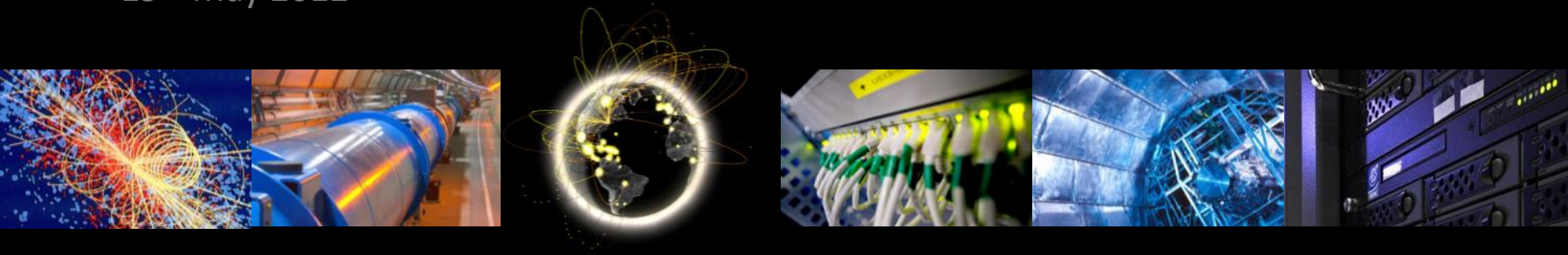
Collaboration Board topics

Ian Bird

CB Meeting; New York

19th May 2012



WLCG
Worldwide LHC Computing Grid

Overview

- New Tier 1s – follow up from last CB
- Helix Nebula (not strictly WLCG)
- WLCG and wider collaboration

New Tier 1s

Background

- In 2011 the first suggestions of potential new Tier 1 sites have been made
- The procedure was initially discussed in the WLCG Collaboration Board in July 2011
- The process is now documented (WLCG-OB-2012-001)
  - And the OB approved this process on 9th March 2012

Process

- Pre-requisite is that any such proposal must be supported by the experiment(s)
- Balance between encouragement of new sites/resources and reaching high standards of existing Tier 1 services
- Process:
  - Prepare with MB a detailed plan that shows how the site will demonstrate required functionality, performance, reliability; timeline and milestones
  - Present plan to OB: OB recommends acceptance (or not)
  - Site can sign MoU as an *Associate* Tier 1
  - MB monitors progress on ramp up, reports to OB
  - When milestones achieved as agreed by MB, final report to OB to recommend full Tier 1 status
    - This should normally take ~1 year

Requirements

- Most elements are described in the MoU addenda
- Candidate site must achieve MoU requirements in terms of:
  - Level and performance of resources – see next
  - Quality and reliability of services:
    - Set of services agreed with the experiments
    - Provide agreed levels of support – as in MoU. Typically on-call support year round
    - Availability and reliability: install agreed sensors, publish to WLCG monthly (as all other sites)
    - Interface to WLCG accounting, provide accounting data to be published monthly
    - Support for Tier 2s – in agreement with experiments. Data source and technical support for Tier 2s

Resources - 1

- Networking:
  - Eventually 10 Gb/s (+ alternate) as part of OPN for T0-T1 or T1-T1
  - Proposal should describe how connectivity provided during prototyping, and the plan to achieve connecting to the OPN
  - Tier 2 connectivity via academic networks. Tier 1 normally expected to have good connectivity. Practical needs to be agreed with experiment(s) – as part of the usage model. Plan should describe this.

Resources - 2

- Must provide tape archive service:
  - Capacity needed for share of raw + other data
  - Must guarantee archive for life of the experiment
  - Specified in MoU
- Must show capability of accepting agreed share of raw data and writing it to tape at agreed rate
  - Plan should detail what this is
- Disk & CPU of significant fraction of experiment requirement
  - Typically ~10% (minimum 5%) of requirements expressed to RRB
  - Must be balanced with adequate internal networking to support expected workloads

New Tier 1s

- At the March OB; KISTI (S. Korea) presented an initial proposal as a Tier 1 for ALICE; the OB accepted KISTI as the first "Associate Tier1"
  - A full plan is now being prepared

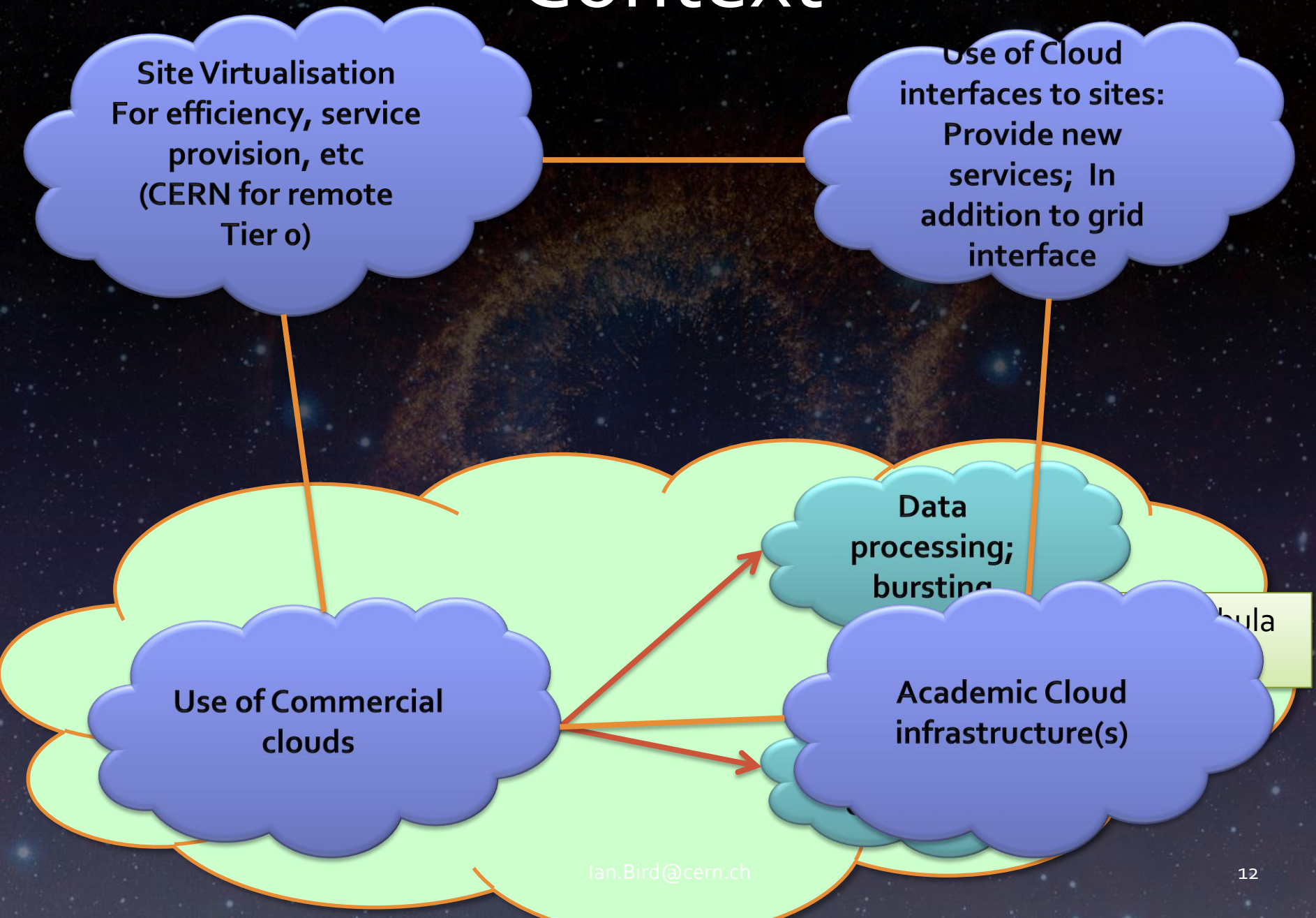
- Also anticipated:
  - Russia has proposed providing Tier 1 for all 4 experiments
  - Discussions with Mexico for ALICE; and India for ALICE and CMS
  - All t.b.c.


HELIX NEBULA

Terminology: EIROForum

- EIROForum
  - Partnership between 8 of Europe's largest inter-governmental scientific research organisations that are responsible for infrastructures and laboratories:
    - CERN, EFDA-JET, EMBL, ESA, ESO, ESRF, European XFEL, and ILL
  - Mission to combine resources, facilities, & expertise of member organisations to support European Science
  - Represents a very large number of European Scientists, and has significant influence on (e.g.) European-level policy and strategy
  - Has a number of collaborative activities/common fields of expertise of which Information Technology is one
  - www.eiroforum.org

Context


Site Virtualisation
For efficiency, service provision, etc
(CERN for remote Tier 0)
Use of Cloud interfaces to sites:
Provide new services; In addition to grid interface
Data processing; bursting
Use of Commercial clouds
Academic Cloud infrastructure(s)

Problem?

- For scientific data processing:
  - HEP data is not restricted in where it can be sent; but other scientific data is (IP, etc.)
    - Concern for ESA, EMBL, etc.
- For other cloud services:
  - Data privacy is a major concern: should not use US companies because of Patriot Act and other concerns about exposure of personnel, financial, and other private data
- Procurement concerns: CERN & others should procure from European (or Member State) sources preferentially
- Thus, cannot just trivially use Amazon, Microsoft, Gmail, DropBox, etc.
  - (the fact that individuals *do* is already a problem!)

Origin of the initiative

- Conceived by ESA as a prospective for providing cloud services to space sector in Europe
- Presented to the IT working group of the EIROforum where other members (CERN, EMBL) joined
- Two workshops held during 2011
  - June: hosted by ESA in Frascati
  - October: hosted by EMBL in Heidelberg

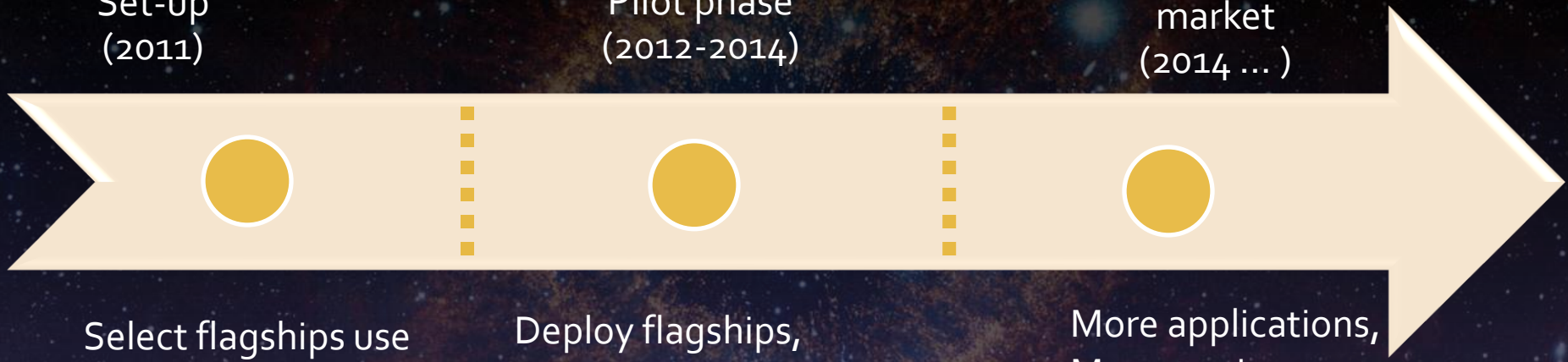
The initiative:

- A strategic plan for a scientific cloud computing infrastructure for Europe
  - Establish a sustainable multi-tenant cloud computing infrastructure in Europe
  - Initially based on the needs for the European Research Area & Space Agencies
  - Based on commercial services from multiple IT industry providers
  - Adhere to internationally recognised policies and quality standards
  - Governance structure involving all stakeholders

Objectives of the initiative

- Set up a cloud computing infrastructure for European Research Area
- Identify and adopt policies for trust, security and privacy on a European-level
- Create a light-weight governance structure involving all stakeholders
- Define a short and medium term funding scheme

Timeline


Set-up
(2011)
Pilot phase
(2012-2014)
Full-scale
cloud service
market
(2014 …)
Select flagships use cases,
identify service providers,
define governance model
Deploy flagships,
Analysis of functionality,
performance & financial
model
More applications,
More services,
More users,
More service providers

Pilot Phase

- Through the pilot phase we expect to explore/push a series of perceived barriers to Cloud adoption:
  - Security: Unknown or low compliance and security standards
  - Reliability: Availability of service for business critical tasks
  - Data privacy: Moving sensitive data to the Cloud
  - Scalability/Elasticity: Will the Cloud scale-up to our needs
  - Network performance: Data transfer bottleneck; QoS
  - Integration: Hybrid systems with in-house/legacy systems
  - Vendor lock-in: Dependency on vendors once data & applications have been transferred to the Cloud
  - Legal concerns: Such as who has legal liability
  - Transparency: Clarity of conditions, terms and pricing

Initial flagships use cases

- Call for proposals
  - Proposals received in format following template agreed by demand and supply side
- Eligibility review of collected proposals (user-side) resulted in 3 recommended flagships
  - CERN: ATLAS High Energy Physics Cloud Use
  - EMBL: Genomic Assembly in the Cloud
  - ESA / CNES / DLR: SuperSites Exploitation Platform
- Flagships sent to supply-side for analysis

Flagship deployments

- Proof of Concept stage within the Pilot Phase started January 2012
- Each flagship will be deployed with a series of providers independently
- Sequence:
  - CERN-ATLAS
  - EMBL
  - ESA
- Expect to have completed initial proof of concept by summer 2012

Flagship use cases
Participating Suppliers

Helix Nebula EC project proposal

Coordination action submitted to INFRA-2012-3.3 in November 2011

- Requested total EC funding €2M

no.	Organisation name	Short name	Country
1 (coord)	European Organization for Nuclear Research	CERN	CH
2	STICHTING EUROPEAN GRID INITIATIVE	EGI.eu	NE
3	European Molecular Biology Laboratory	EMBL	DE
4	ATOS ORIGIN NEDERLAND	Atos	NE
5	T-Systems International GMBH	T-Systems	DE
6	CLOUDSIGMA AG	CloudSigma	CH
7	SAP AG	SAP	DE
8	Logica Deutschland GmbH & Co KG	Logica	DE
9	CONSIGLIO NAZIONALE DELLE RICERCHE	CNR	IT
10	Cloud Security Alliance EMEA	CSA	UK

Role of Helix Nebula: The Science Cloud

Vision of a unified cloud-based infrastructure for the ERA based on Public/Private Partnership, 4 goals building on the collective experience of all involved.

- ***Goal One*** : Establish HELIX NEBULA – *the Science Cloud* – as a cloud computing infrastructure addressing the needs of the ERA and capable of serving as a platform for innovation and evolution of the overall e-infrastructure.
- ***Goal Two:*** Identify and adopt suitable policies for trust, security and privacy on a European-level
- ***Goal Three:*** Create a light-weight governance structure that involves all the stakeholders - and which can evolve over time as the infrastructure, services and user-base grows.
- ***Goal Four:*** Define a funding scheme involving all the stake-holder groups (service suppliers, users, EC and national funding agencies) for PPP to implement a Cloud Computing Infrastructure that delivers a sustainable and profitable business environment adhering to European- level policies.

Specific outcomes

- Develop **strategies for extremely large or highly distributed and heterogeneous scientific data** (including service architectures, applications and standardisation) in order to manage the upcoming data deluge
- Analyse and promote trust building towards open scientific data e-Infrastructures covering **organisational, operational, legal and technological aspects,** including authentication, authorisation and accounting (AAA)
- Develop strategies and establish structures aiming at **co-ordination between e-infrastructure operators**
- Create frameworks, including **business models for supporting Open Science and cloud infrastructures based on PPP,** useful for procurement of computing services suitable for e- Science

Scientific Flagships

- CERN LHC (ATLAS):
  - High Throughput Computing and large scale data movement
- EMBL:
  - Novel *de novo* genomic assembly techniques
- ESA:
  - Integrated access to data held in existing Earth Observation "Super Sites"
- Each flagship brings out very different features and requirements and exercises different aspects of a cloud offering

ATLAS use case

- Simulations (~no input) with stage out to:
  - Traditional grid storage vs
  - Long term cloud storage
- Data processing (== "Tier 1")
  - This implies large scale data import and export to/from the cloud resource
- Distributed analysis (== "Tier 2")
  - Data accessed remotely (located at grid sites), or
  - Data located at the cloud resource (or another?)
- Bursting for urgent tasks
  - Centrally managed: urgent processing
  - Regionally managed: urgent local analysis needs
- All experiences immediately transferable to other LHC (& HEP) experiments

Immediate and longer term goals for CERN

- Determine costs of commercial cloud resources from various sources
  - Compute resources
  - Network transfers into and out of cloud
  - Short and long term data storage in the cloud
- Develop understanding of appropriate SLA's
  - How can they be broadly applicable to LHC or HEP
- Understand policy and legal constraints; e.g. in moving scientific data to commercial resources
- Performance and reliability – compared to WLCG baseline
- Use of standards (interfaces, etc.) & interoperability between providers
- Can CERN transparently offload work to a cloud resource
  - Which type of work makes sense?
- Long term: Can we use commercial services as a significant fraction of overall resources available to CERN experiments?
  - At which point is it economic/practical to rely on 3rd party providers?

Summary

- The objective of this initiative is to establish a sustainable cloud computing infrastructure for the European Research Area based on commercially provided services
- It is a collaborative initiative bringing together all the stakeholders to establish a public-private partnership
- Interoperability with existing e-infrastructures is a goal of the initiative
- It has commitments from the IT industry and user organisations – flagship deployment started Jan 2012
- 3 initial flagship use cases identified
- Framework collaboration EC project to start summer 2012

WLCG & HEP

- "WLCG" (and OSG, EGI, NDGF, etc.) used already by other HEP experiments and others
- Suggestions by SuperBelle, and others that they could "join" WLCG and benefit from not just the infrastructure and technology but also support, operations, etc.
- Other initiatives looking at broader use of e-infrastructures
- How should WLCG collaboration position itself?
  - We are only mandated for LHC today
  - How is it different from EGI, OSG, etc?
    - It is worldwide, collaborates across infrastructures