



Grid Tools & Services

Rob Gardner
University of Chicago

ATLAS Software Meeting
August 27-29, 2003
BNL

Outline

- Brief overview of Grid projects
 - US and LCG
- Ongoing US ATLAS development efforts
 - Grid data management, packaging, virtual data tools
- ATLAS PreDC2 Exercise and Grid3
 - Joint project with CMS
 - Fall demonstrator exercise



US ATLAS and Grid Projects

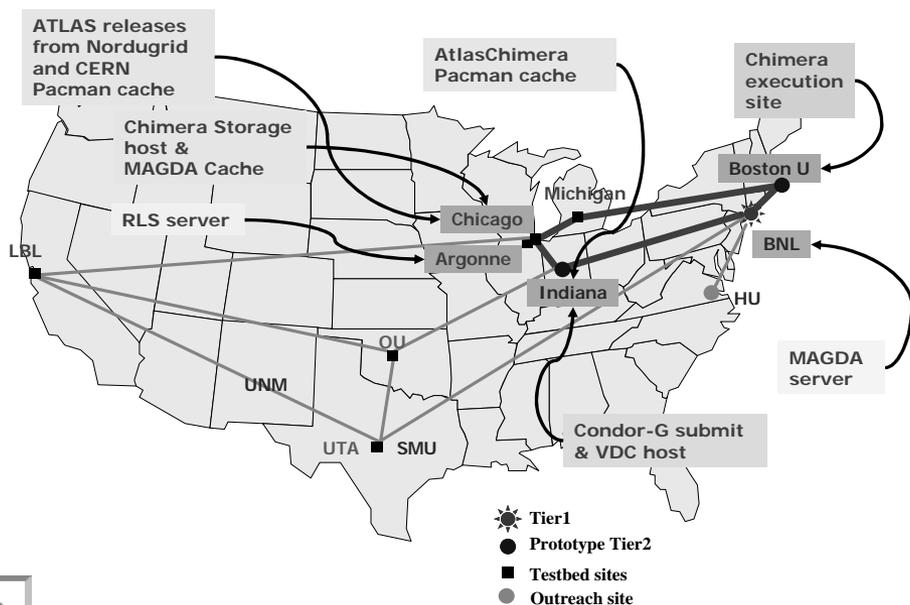


- Particle Physics Data Grid (PPDG)
 - End to end applications and services
 - ATLAS activity: distributed data storage on the grid (Magda), EDG interoperability, Chimera end-to-end applications
- GriPhyN
 - CS and Physics collaboration to develop virtual data concept
 - Physics: ATLAS, CMS, LIGO, Sloan Digital Sky Survey
 - CS: build on existing technologies such as Globus, Condor, fault tolerance, storage management, plus new computer science research
- iVDGL
 - International Virtual Data Grid Laboratory
 - Platform on which to design, implement, integrate and test grid software
 - Infrastructure for ATLAS and CMS prototype Tier2 centers
 - Forum for grid interoperability – collaborate with EU DataGrid, DataTag



US LHC Testbeds

- US ATLAS and US CMS each have active testbed groups building production services on VDT based grids



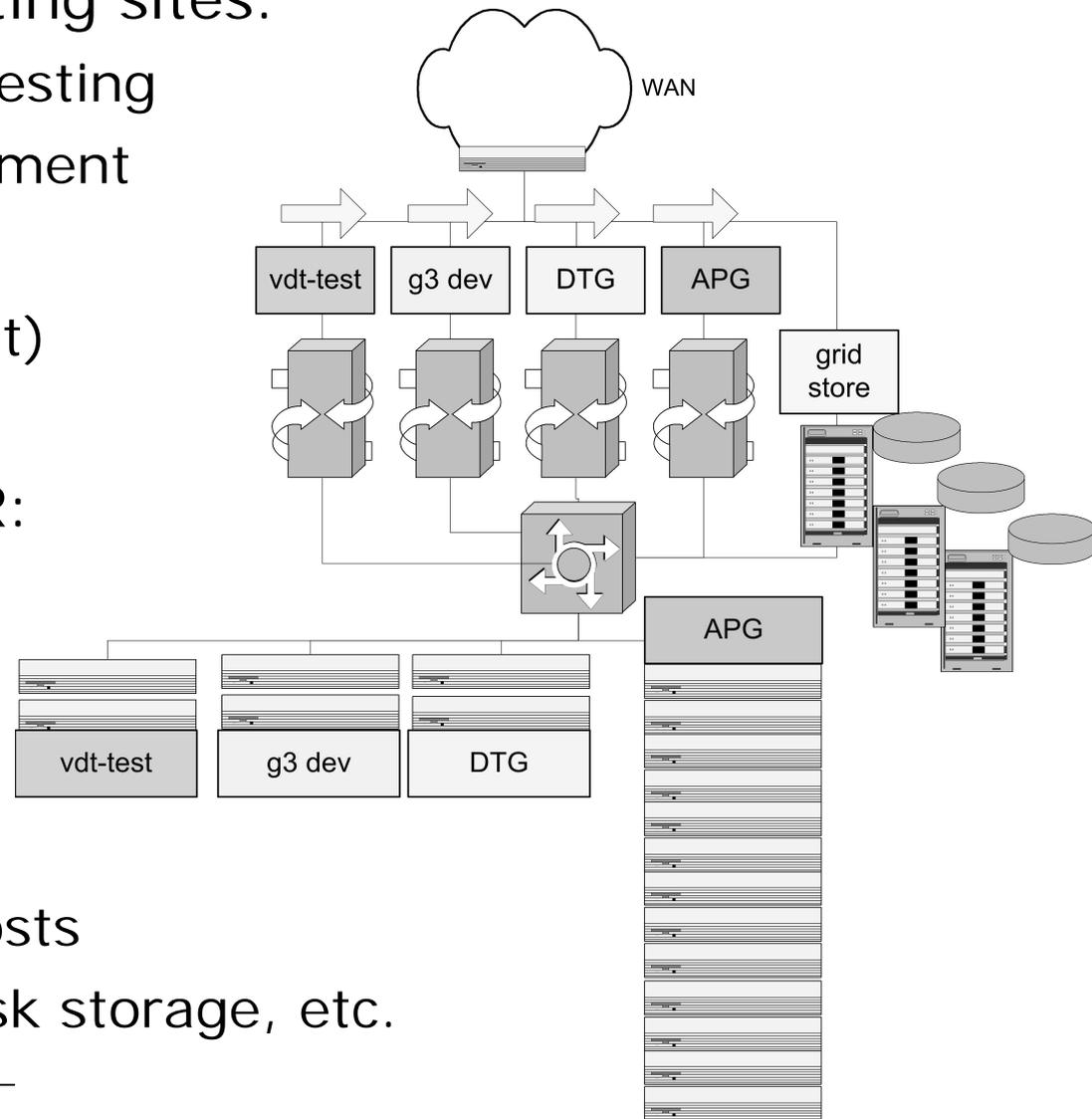
ATLAS

DTG – Development Test Grid
 US ATLAS Production Testbed
 VDT Testgrid



Example Site Configuration

- For component testing sites:
 - Pre-release VDT-testing
 - Pre-Grid3 environment
- US ATLAS Grids
 - DTG (development)
 - APG (production)
- Grid servers, L to R:
 - stability increases
 - more resources
 - cycle less often
- Other machines
 - Client (submit) hosts
 - RLS, VDC, grid disk storage, etc.



VDT I

- VDT (Virtual Data Toolkit) is the baseline middleware package for US LHC and LCG projects
- Delivered by the GriPhyN and iVDGL CS-Physics grid projects, with contributions from several technology providers
- Core Grid services based on Globus and Condor
 - Globus security infrastructure (GSI)
 - Globus Information Services (MDS)
 - Globus Resource Management (GRAM)
 - Globus Replica Location Service (RLS)
 - Condor Execution
 - > CondorG (Condor to Globus)
 - > Directed Acyclic Graph Manager (DAGMAN)



VDT II

- Virtual Data Software
 - Virtual Data Language, abstract planning, and Catalog (Chimera)
 - Concrete Planning (Pegasus)
 - Scheduling Framework (Sphynx)
- Virtual Organization Management Software
 - > VOMS and extensions (VOX)
 - > Grid User Management System (GUMS)
- Packaging software (Pacman)
- Monitoring Software – Mona Lisa, NetLogger
- VDT Testers Group provides pre-release grid testing



LCG – Goals (I.Bird)

- The goal of the LCG project is to prototype and deploy the computing environment for the LHC experiments
- Two phases:
 - **Phase 1: 2002 – 2005**
 - Build a service prototype, based on existing grid middleware
 - Gain experience in running a production grid service
 - Produce the TDR for the final system
 - **Phase 2: 2006 – 2008**
 - Build and commission the initial LHC computing environment



LCG is not a development project – it relies on other grid projects for grid middleware development and support



LCG Regional Centres

Centres taking part in the LCG prototype service : 2003 – 2005

Tier 0

- CERN

Tier 1 Centres

- Brookhaven National Lab
- CNAF Bologna
- Fermilab
- FZK Karlsruhe
- IN2P3 Lyon
- Rutherford Appleton Lab (UK)
- University of Tokyo
- CERN

Other Centres

- Academia Sinica (Taipei)
- Barcelona
- Caltech
- GSI Darmstadt
- Italian Tier 2s(Torino, Milano, Legnaro)
- Manno (Switzerland)
- Moscow State University
- NIKHEF Amsterdam
- Ohio Supercomputing Centre
- Sweden (NordusGrid)
- Tata Institute (India)
- Triumf (Canada)
- UCSD
- UK Tier 2s
- University of Florida–Gainesville
- University of Prague
-

Confirmed Resources: http://cern.ch/lcg/peb/rc_resources



LCG Resource Commitments – 1Q04

	<i>CPU (kSI2K)</i>	<i>Disk TB</i>	<i>Support FTE</i>	<i>Tape TB</i>
<i>CERN</i>	700	160	10.0	1000
<i>Czech Republic</i>	60	5	2.5	5
<i>France</i>	420	81	10.2	540
<i>Germany</i>	207	40	9.0	62
<i>Holland</i>	124	3	4.0	12
<i>Italy</i>	507	60	16.0	100
<i>Japan</i>	220	45	5.0	100
<i>Poland</i>	86	9	5.0	28
<i>Russia</i>	120	30	10.0	40
<i>Taiwan</i>	220	30	4.0	120
<i>Spain</i>	150	30	4.0	100
<i>Sweden</i>	179	40	2.0	40
<i>Switzerland</i>	26	5	2.0	40
<i>UK</i>	1780	455	24.0	300
<i>USA</i>	801	176	15.5	1741
Total	5600	1169	123.2	4228



Deployment Goals for LCG-1

- Production service for Data Challenges in 2H03 & 2004
 - Initially focused on batch production work
 - But '04 data challenges have (as yet undefined) interactive analysis
- Experience in close collaboration between the Regional Centres
 - Must have wide enough participation to understand the issues,
- Learn how to maintain and operate a global grid
- Focus on a production-quality service
 - Robustness, fault-tolerance, predictability, and supportability take precedence; additional functionality gets prioritized
- LCG should be integrated into the sites' physics computing services – should not be something apart
 - This requires coordination between participating sites in:
 - > Policies and collaborative agreements
 - > Resource planning and scheduling
 - > Operations and Support



Elements of a Production LCG Service

- Middleware:
 - Testing and certification
 - Packaging, configuration, distribution and site validation
 - Support – problem determination and resolution; feedback to middleware developers
- Operations:
 - Grid infrastructure services
 - Site fabrics run as production services
 - Operations centres – trouble and performance monitoring, problem resolution – 24x7 globally
- Support:
 - Experiment integration – ensure optimal use of system
 - User support – call centres/helpdesk – global coverage; documentation; training



LCG-1 Components, at a glance

- Scope, from LCG Working Group 1 report:
 - *“The LCG-1 prototype Grid has as its primary target support of the production processing needs of the experiments for their data challenges. It does not aim to provide a convenient or seamless environment for end-user analysis.”*
- Components (VDT and EDG based)
 - User Tools and Portals: not in scope
 - Information Services: GLUE compliant
 - AAA and VO: EDG VOMS
 - Data management: EDG RLS
 - Virtual Data Management: not covered
 - Job Management and Scheduling: EDG WP1 RB
 - Application Services and Higher Level Tools: none
 - Packaging and Configuration: GDB/WP4 and GLUE groups looking at this (lcfgng, pacman,...)
 - Monitoring: Nagios, Ganglia considered
- **Testbed deployments underway at BNL and FNAL**



Tools and Services

MAGDA
Pacman
Chimera
DIAL
GRAPPA
GANGA
GRAT
Ganglia
GUMS
MonaLisa
VOMS
VOX



MAGDA: Manager for grid-based data

- MAnager for Grid-based DAta prototype.
- Designed for rapid development of components to support users quickly, with components later replaced by Grid Toolkit elements.
 - Deploy as an evolving production tool.
- Designed for 'managed production' *and* 'chaotic end-user' usage.
- Adopted by ATLAS for Data Challenge 0 and 1.

Info: <http://www.atlasgrid.bnl.gov/magda/info>

The system: <http://www.atlasgrid.bnl.gov/magda/dyShowMain.pl>

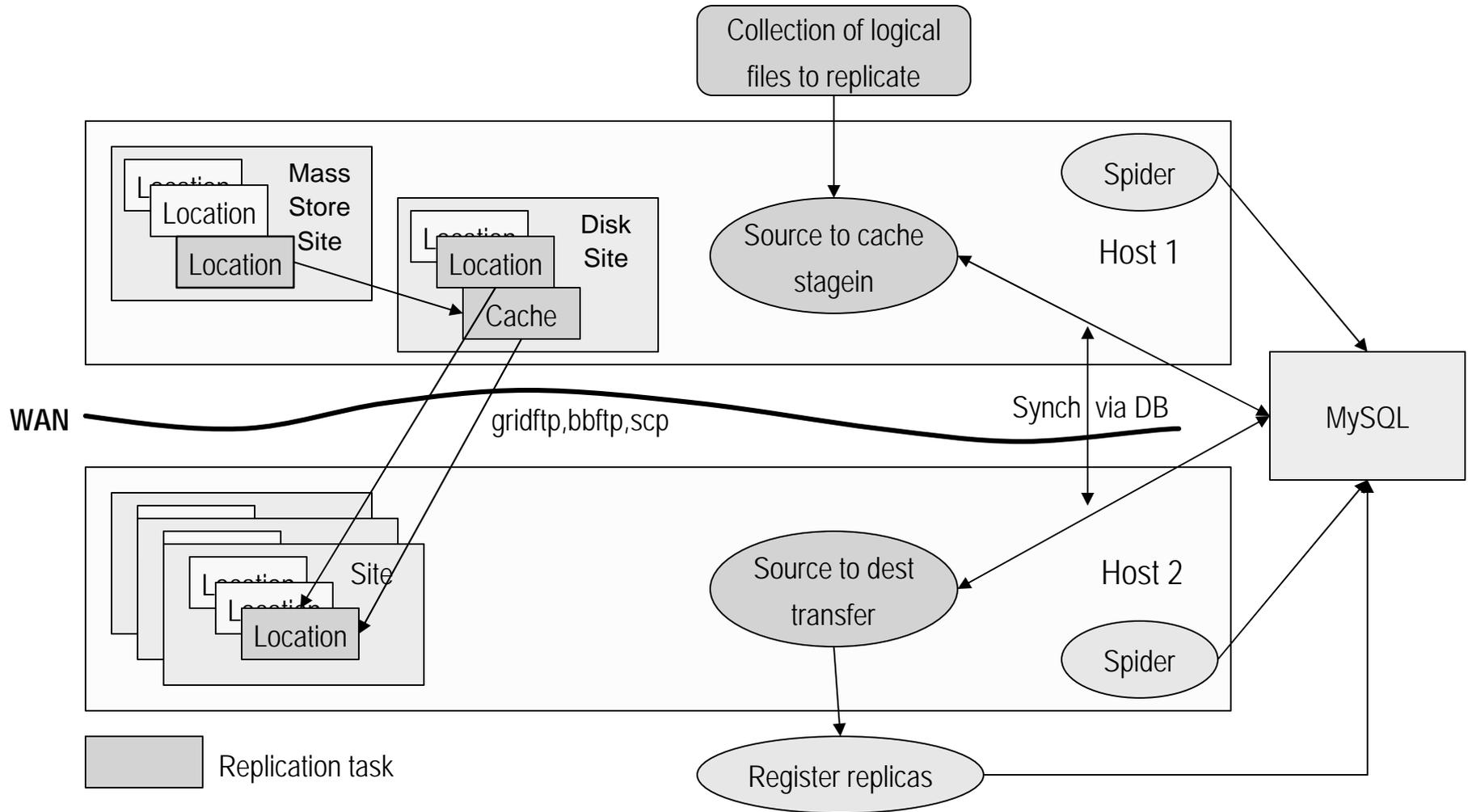


Architecture and Schema

- MySQL database at the core of the system.
 - DB interaction via perl, C++, java, cgi (perl) scripts
 - C++ and Java APIs autogenerated off the MySQL DB schema
- User interaction via web interface and command line.
- Principal components:
 - **File catalog** covering any file types
 - **Data repositories** organized into **sites**, each with its **locations**
 - Computers with repository access: a **host** can access a set of **sites**
 - Logical files can optionally be organized into **collections**
 - **Replication** operations organized into **tasks**



Magda Architecture



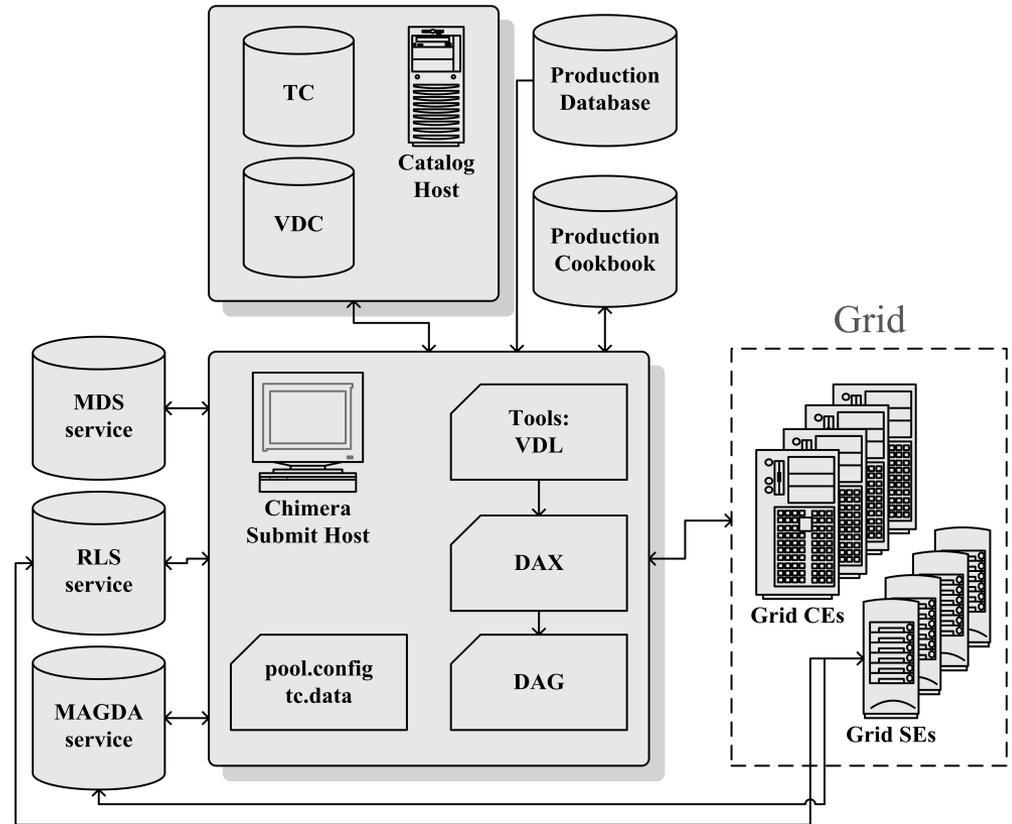
Pacman

- Package and installation manager for the Grid
- Adopted by VDT, US LHC Grid projects, NPACI, ...
- New features expected soon:
 - Dependent and remote installations; can handle Grid elements or farm components behind a gatekeeper
 - Can resolve against snapshots rather than the caches – very useful for virtual data
 - Robustness: integrity-checking, repair, uninstall and reinstall features, better handling of circular dependencies
 - Advanced warning of installation problems
 - Consistency checking between packages
 - Installation from cache subdirectories
 - Setup scripts generated for installation subpackages
 - Better version handling
 - More info: <http://physics.bu.edu/~youssef/>



Grid Component Environment

- Components used from several sources:
 - ATLAS software
 - VDT middleware
 - Services: RLS, Magda, Production database, Production cookbook
 - RLS: primary at BNL, secondary at UC
- Client host package:
 - GCL: GCE-Client
 - > VDT Client, Chimera/Pegasus, other client tools (Magda, Cookbook, ...)



As used in DC1 Reconstruction

- Athena-based reconstruction
- Installed ATLAS releases 6.0.2+ (Pacman cache) on select US ATLAS testbed sites
- 2x520 partitions of DataSet 2001 (lumi10) have been reconstructed at JAZZ-cluster (Argonne), LBNL, IU and BU, BNL (test)
- 2x520 Chimera derivations, ~200,000 events reconstructed
- Submit hosts - LBNL; others: Argonne, UC, IU
- RLS-servers at the University of Chicago and BNL
- Storage host and Magda cache at BNL
- Group-level Magda registration of output
- Output transferred to BNL and CERN/Castor



DC2 ... Setting the requirements

- DC2: Q4/2003 – Q2/2004
- Goals
 - Full deployment of Event Data Model & Detector Description
 - Geant4 becomes the main simulation engine
 - Pile-up in Athena
 - Test the calibration and alignment procedures
 - Use LCG common software
 - Use widely GRID middleware
 - Perform large scale physics analysis
 - Further tests of the computing model
- Scale
 - As for DC1: $\sim 10^7$ fully simulated events (pile-up too)
- ATLAS-wide production framework needed

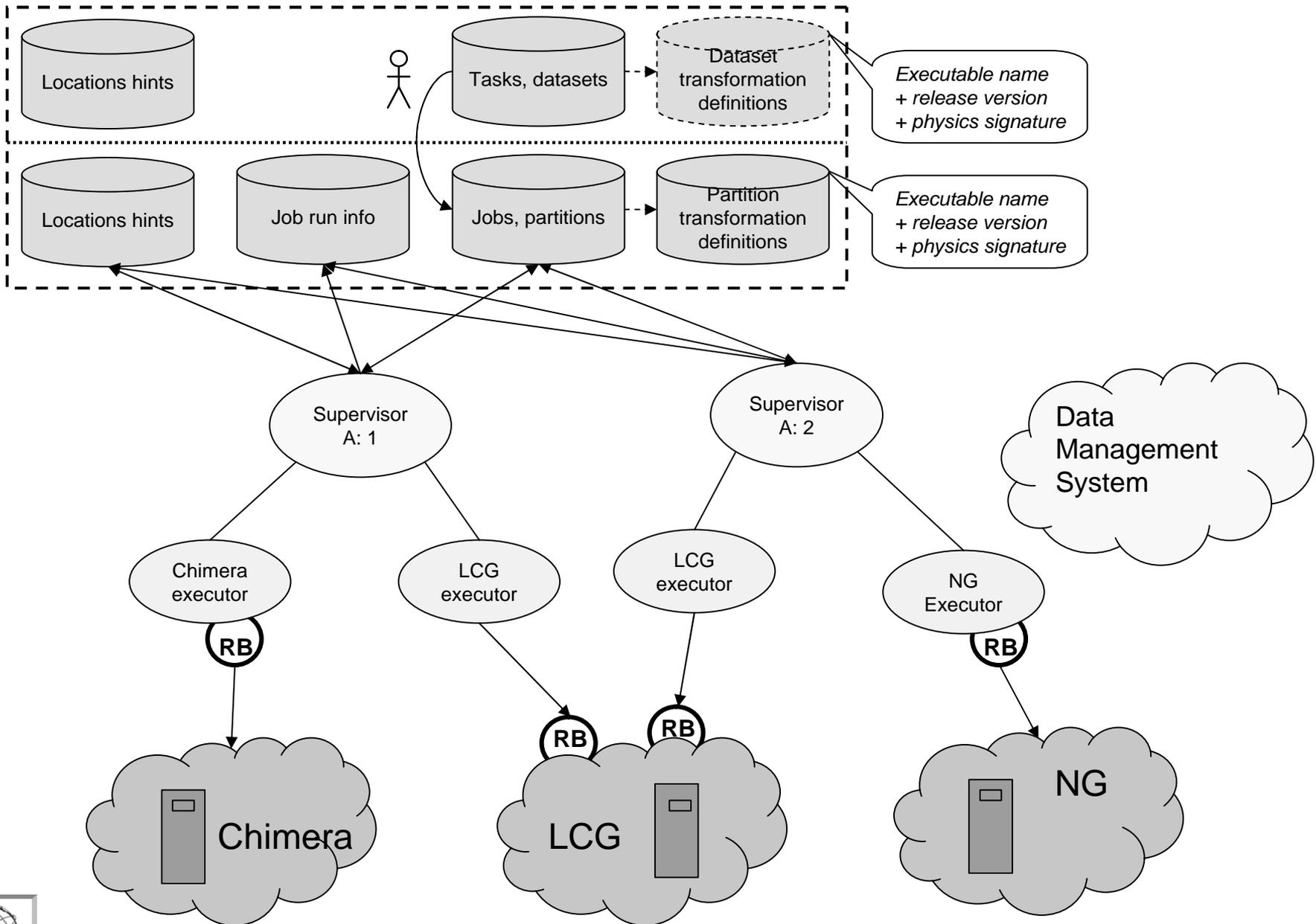


Broader Considerations

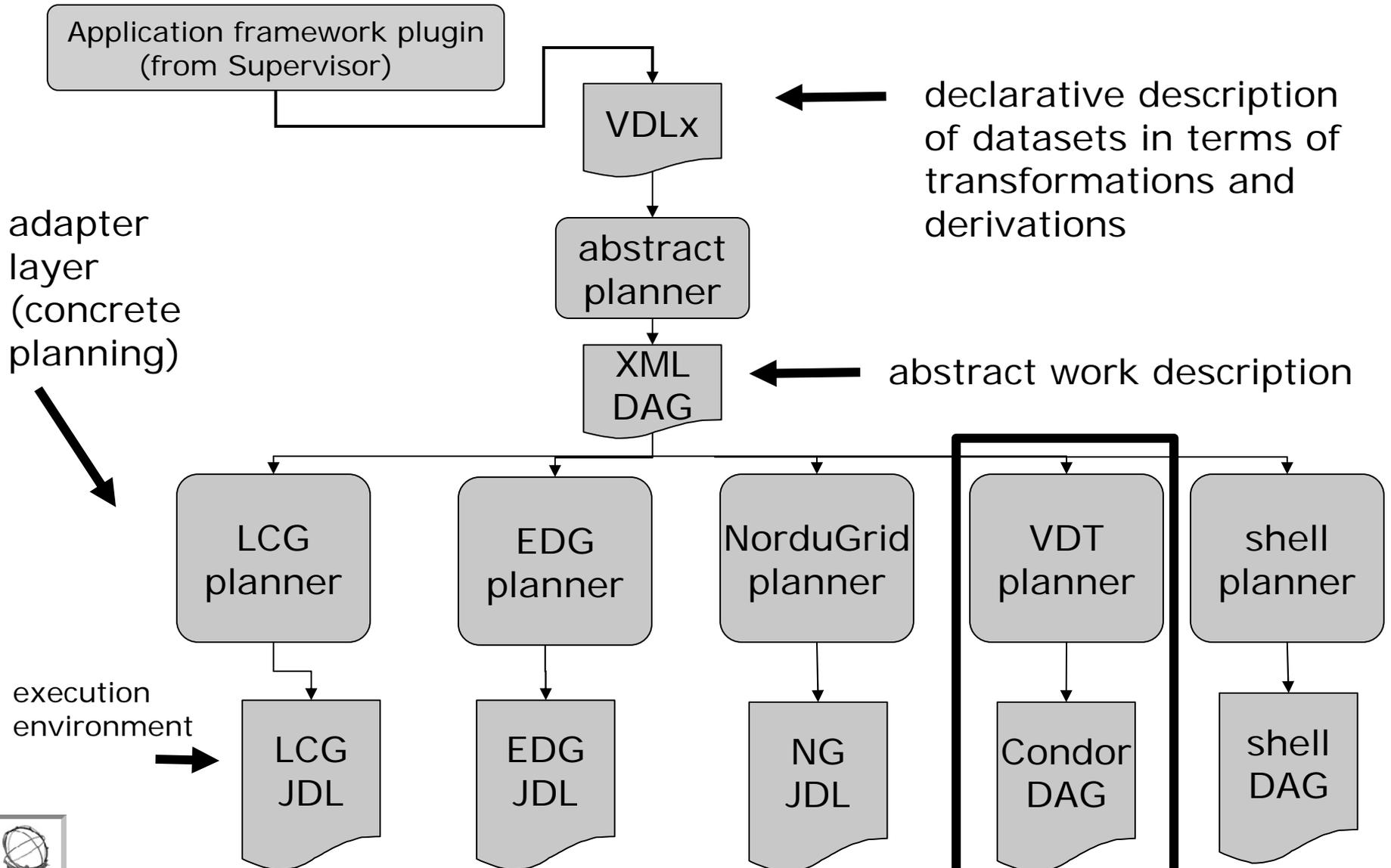
- Need a level of task management above the services which LCG-1 will provide
- Need to connect to ATLAS development efforts on interfaces and catalogs
- Reuse components for analysis
- Would like to use several grid middleware services besides LCG-1:
 - EDG, VDT, NorduGrid, ..

- ➔ Suggests flexible applications and user interfaces on the front-end, multiple grid execution environments on the backend
- ➔ Need to find a point of integration in the middle





Example Chimera Executor



ATLAS PreDC2 and Grid3



Grid3 Background

- A grid environment project among:
 - US ATLAS and US CMS Software and Computing projects
 - iVDGL, PPDG, and GriPhyN grid projects
 - SDSS and LIGO computing projects
 - > These are the Stakeholders of the project
- This builds on successful testbed efforts, data challenges, interoperability work (WorldGrid demonstration)
 - Will incorporate new ideas in VO project management being developed now, and others
 - Project explicitly includes production and analysis
- Plan endorsed by Stakeholders in iVDGL Steering meeting
 - <http://www.ivdgl.org/planning/2003-06-steering/>



What is Grid3?

- Grid3 is a project to build a grid environment to:
 - Provide the next phase of the iVDGL Laboratory
 - Provide the infrastructure and services need for LHC production and analysis applications running at scale in a common grid environment
 - Provide a platform for computer science technology demonstrators
 - Provide a common grid environment for LIGO and SDSS applications



Grid3 Strategies

- Develop, integrate and deploy a functional grid across LHC institutions, extending to non-LHC institutions and to international sites, working closely together with the existing efforts.
- Demonstrate the functionalities and capabilities of this Grid, according to some well-defined metrics.
- The Grid3 project major demonstration milestone coincides with the SC2003 conference
- Work within the existing VDT-based **US LHC Grid** efforts, towards a **functional** and capable Grid.
- The Grid3 project itself would end at the end of 2003.
 - Assuming the metrics would be met.
 - With the expectation that the Grid infrastructure would continue to be available for applications and scheduled iVDGL lab work,
 - A follow-up work plan will be defined, well in advance of the project end.



ATLAS PreDC2 and Grid3

- Monte Carlo production
 - if possible extending to non-U.S. sites, using Chimera-based tools
- Collect and archive MC data at BNL
 - MAGDA, RLS/RLI/RLRCs involved
- Push MC data files to CERN
 - MAGDA
- Reconstruction at CERN
 - using LHC/EDG components
- MAGDA "spider" finds new reconstruction output in RLI and copies them to BNL
 - may require interfacing to EDG RLS
- Data reduction at BNL, creating "collections" or "datasets" and skimming out n-tuples
 - DIAL, maybe Chimera
- Analysis: Distributed analysis of collections and datasets.
 - DIAL, GANGA



Grid3 Core Components

- A grid architecture consisting of facilities (eg., execution and storage sites), services (eg. a prototype operations center, information system), and applications.
- The middleware will be based on VDT 1.1.10 or greater, and will use components from other providers as appropriate (eg. Ganglia, LCG).
- Applications
- Other services (eg. RLS)



Core Components, 2

- An information service based on MDS will be deployed.
- A simple monitoring service based on Ganglia, version 2.5.3 or greater
 - (supporting hierarchy, grid-level collections), with collectors at one or more of the operations centers. Additionally, Ganglia information providers to MDS may be deployed if required, and native Ganglia installations may report to other collectors corresponding to other (logical) grids. An example is a collector at Fermilab providing a Ganglia grid-view of the US CMS managed resources.
 - Nagios for alarms and status monitoring should be investigated
- Other monitoring systems, such as MonaLisa or a workflow monitoring system, will be deployed.



Core Components, 3

- A consistent method or set of recommendations for packaging, installing, configuring, and creating run-time environments for applications, subject to review.
- Use of the WorldGrid project mechanism
- One or more VO management mechanisms for authentication and authorization will be used.
 - The VOMS server method as developed by the VOX project.
 - The WorldGrid project method as developed by Pacman.
 - A fallback solution is to use LDAP VO servers, one for each VO containing the DN's of the expected application users.

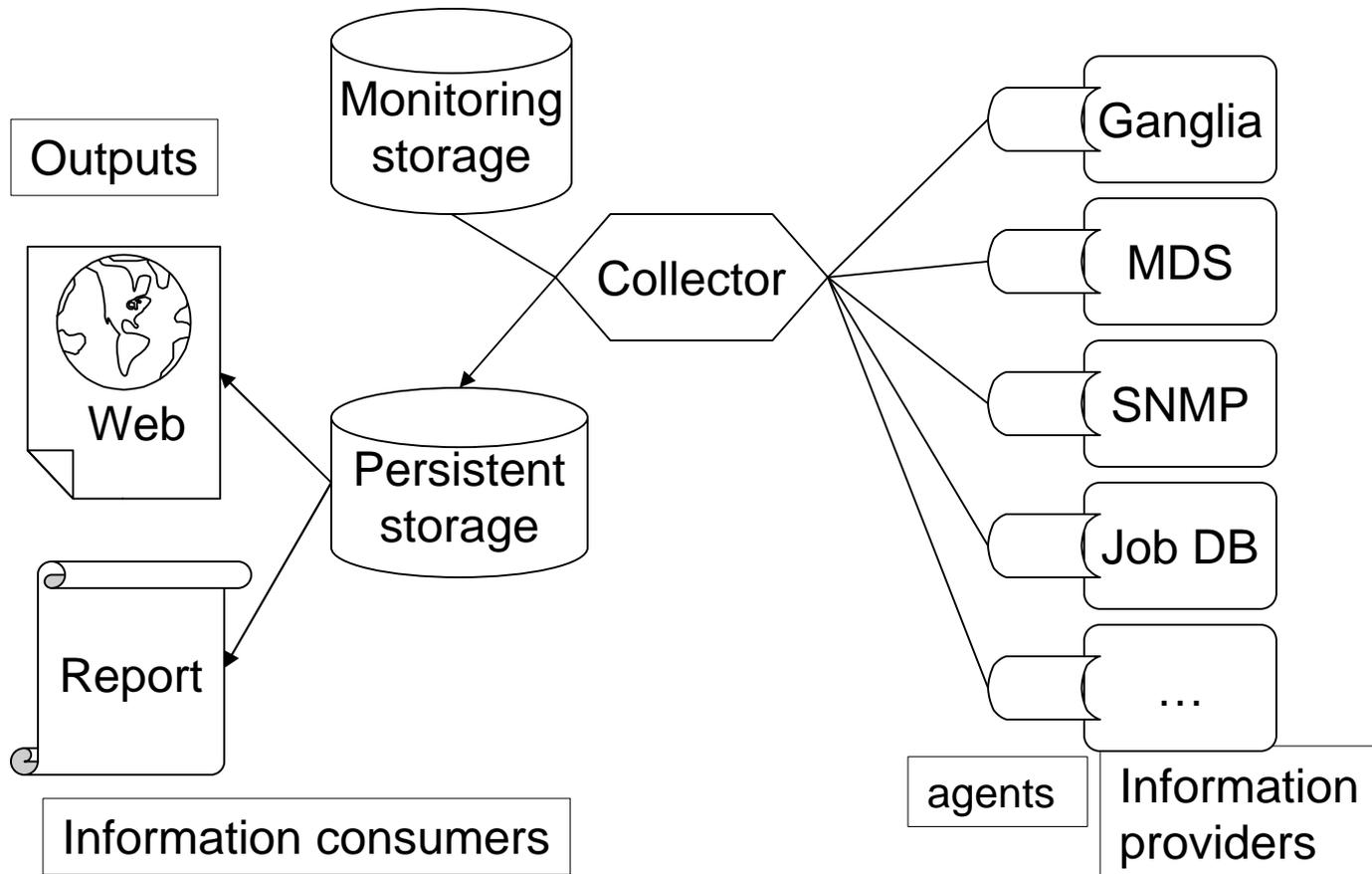


Metrics

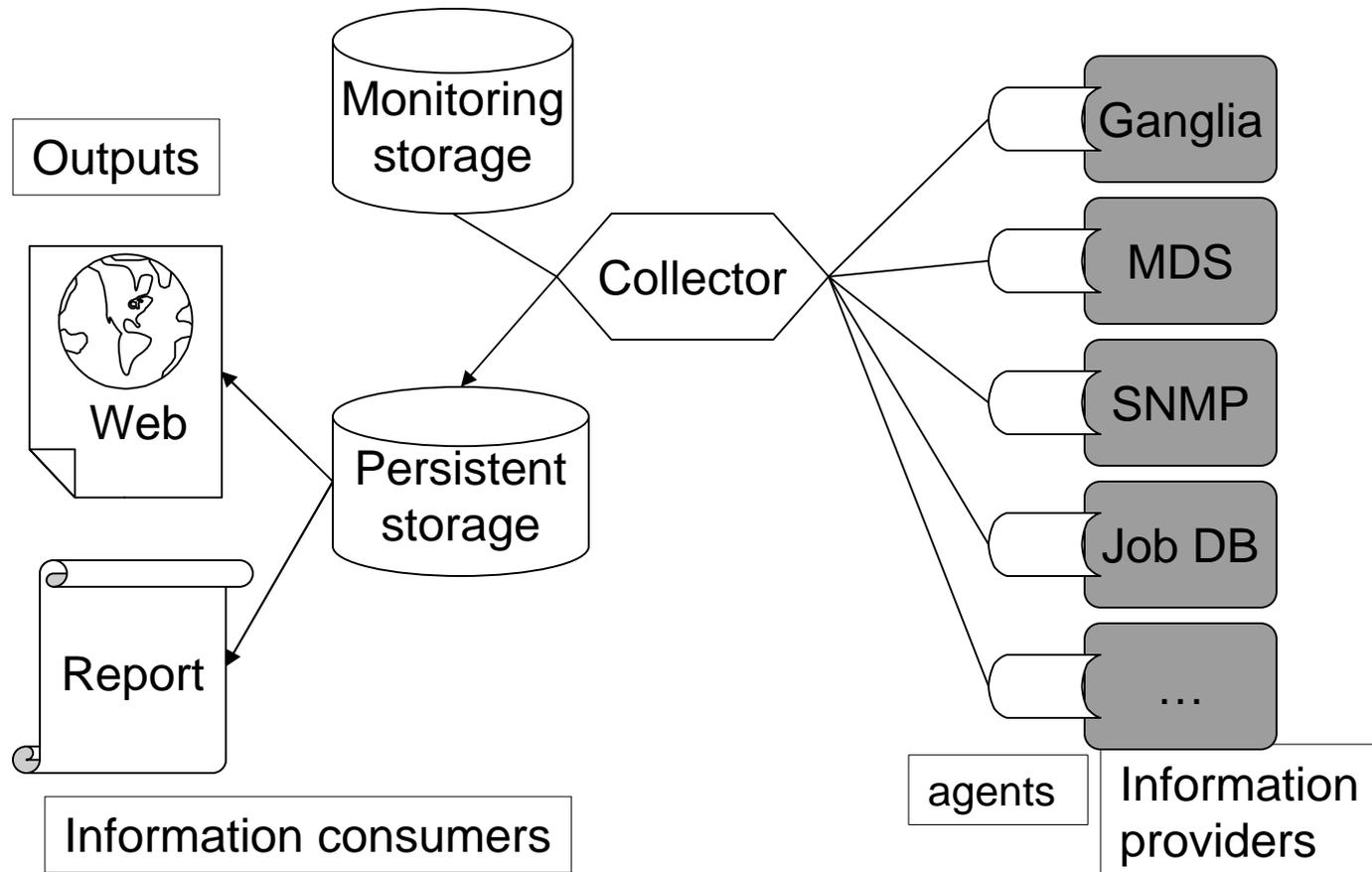
- In order to characterize the computing activity and gauge the resulting performance, we define additional goals of the project to include metrics and feature demonstrations. These will need to be discussed at the beginning of the project. We expect the project to include deployment of mechanisms to collect, aggregate, and report measurements.
- Examples:
 - Data transferred per day > 1 TB
 - Number of concurrent jobs > 100
 - Number of users > 10
 - Number of different applications > 4
 - Number of sites running multiple applications > 10
 - Rate of Faults/Crashes < 1/hour
 - Operational Support Load of full demonstrator < 2 FTEs



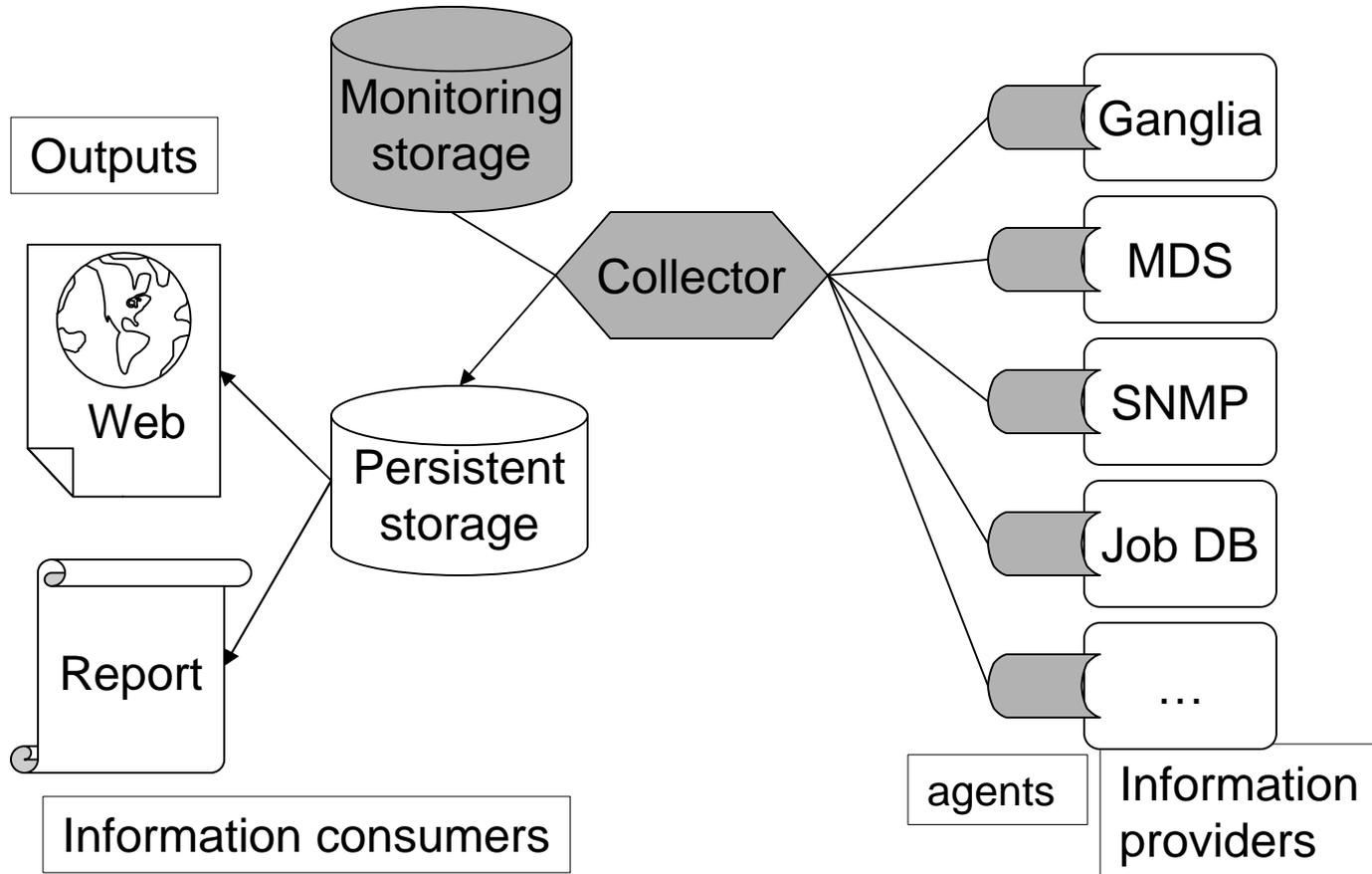
Monitoring framework



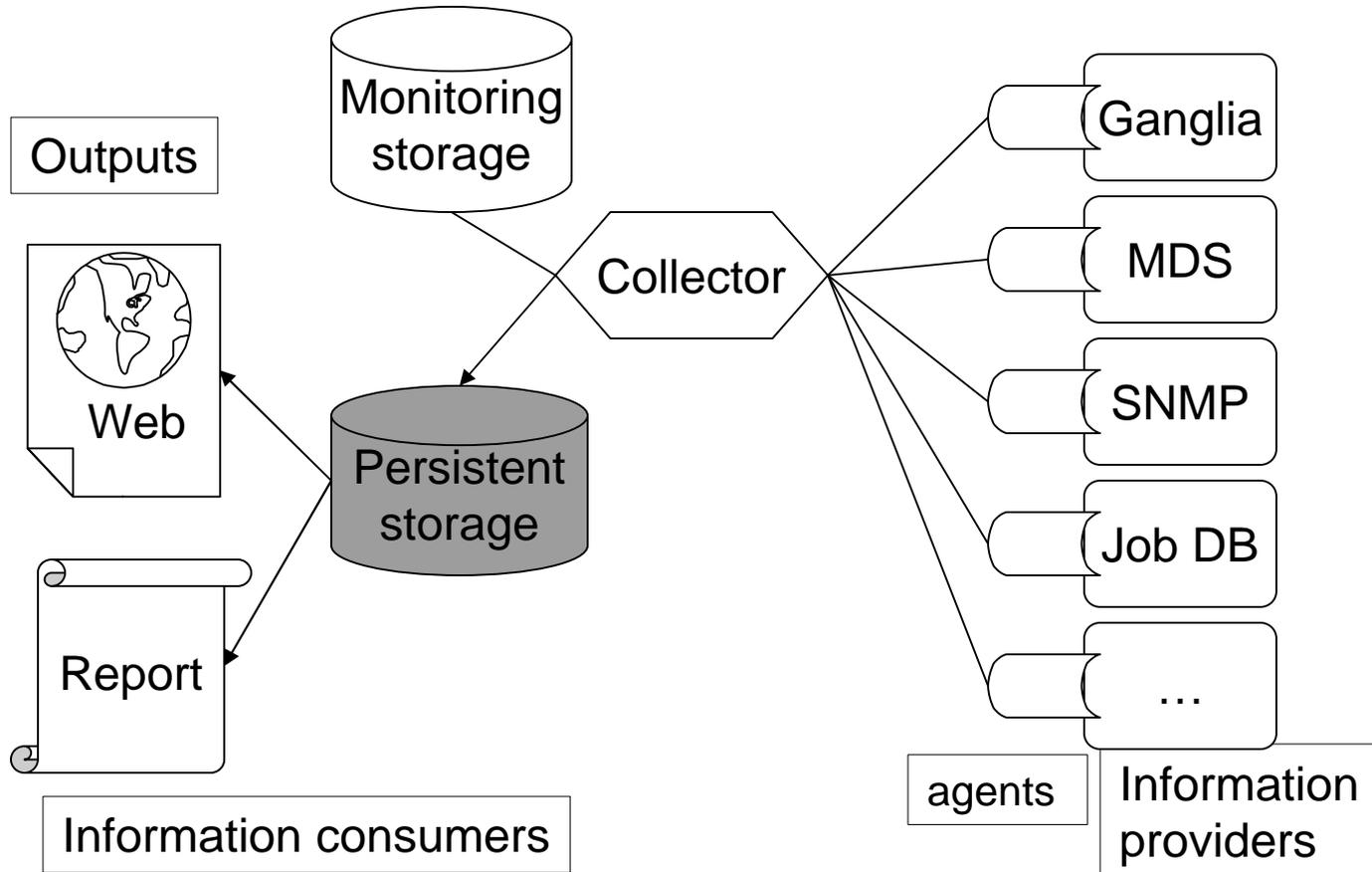
Information providers (on CE/WN)



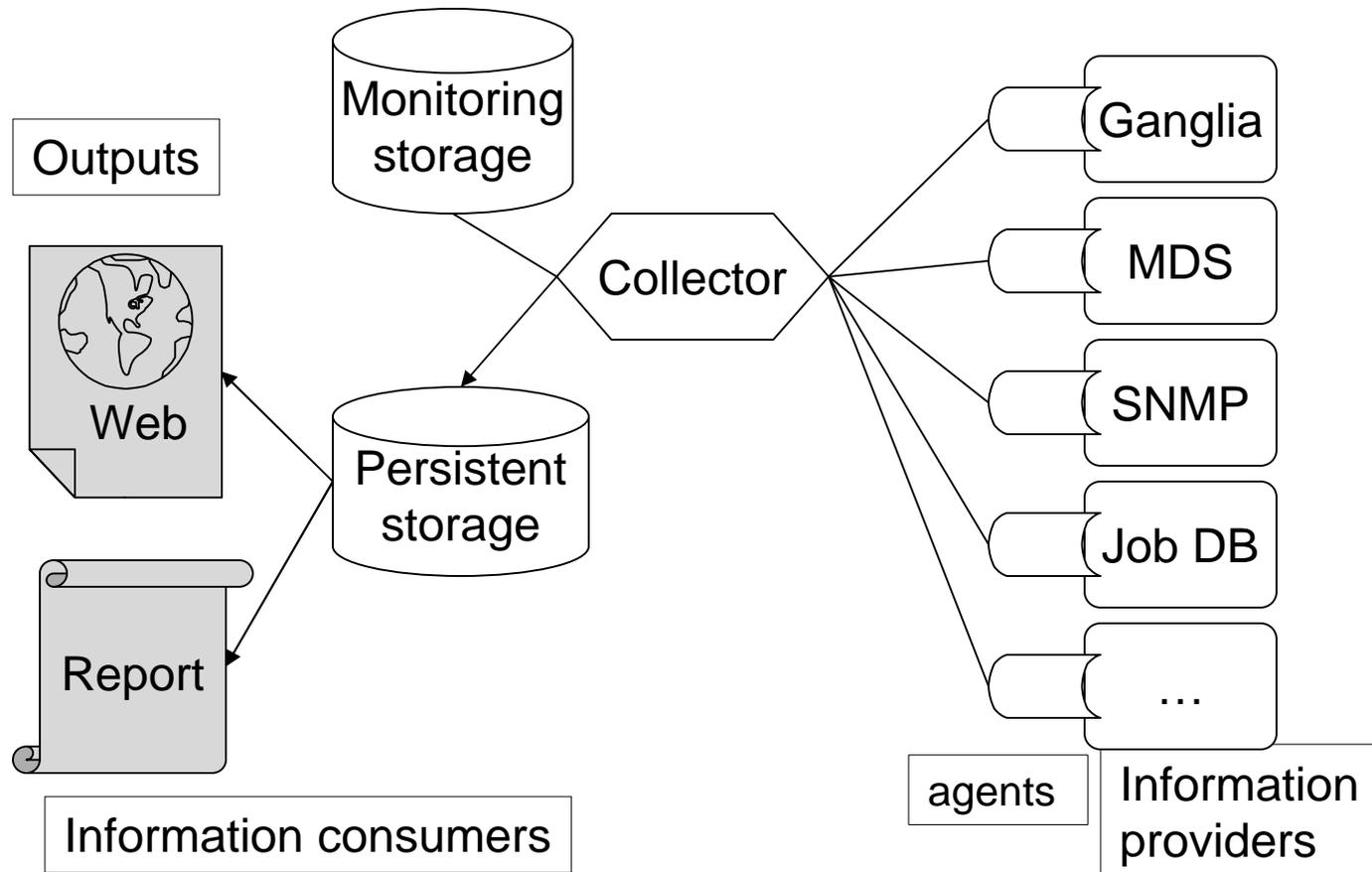
MonALISA



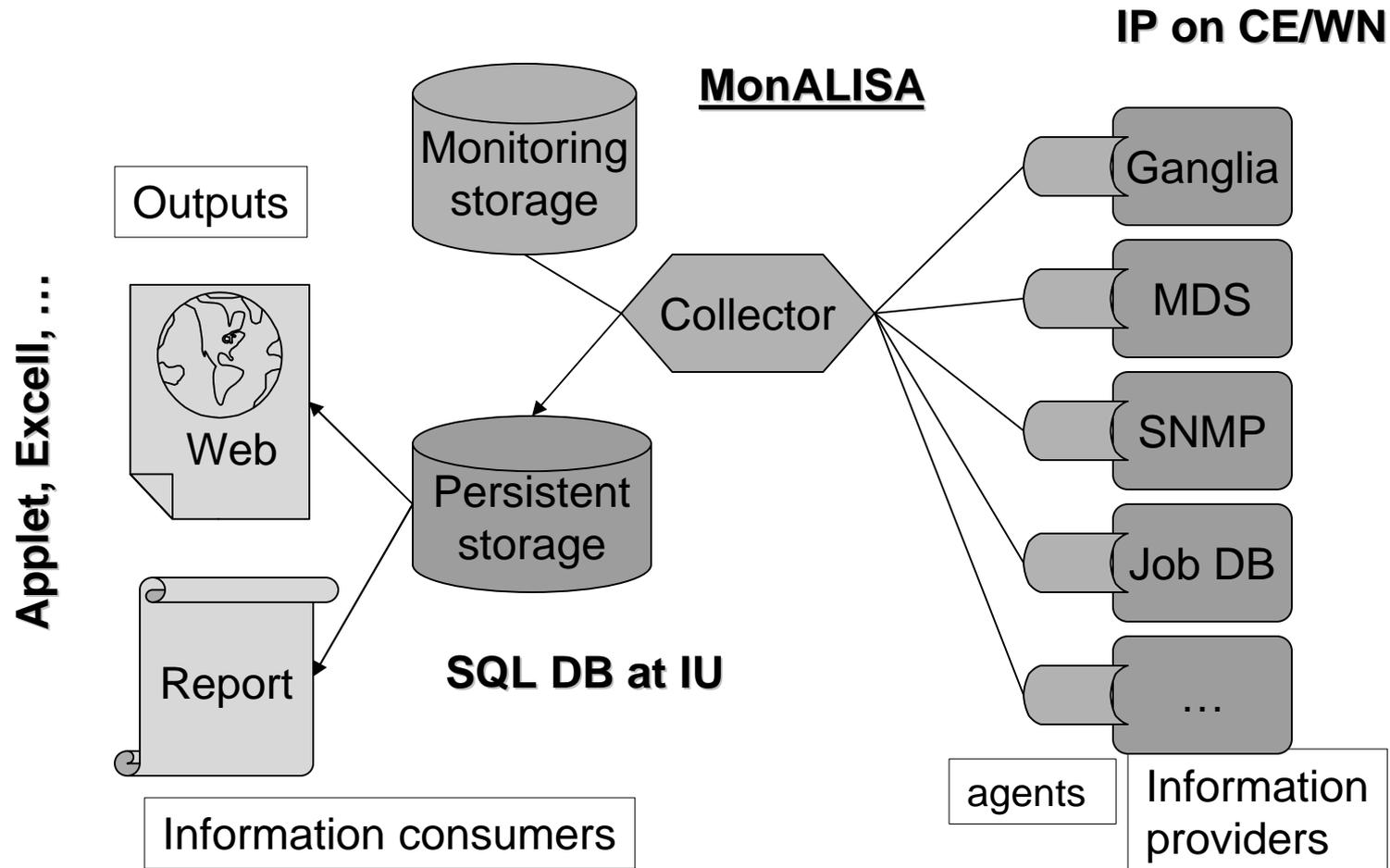
SQL DataBase at IU



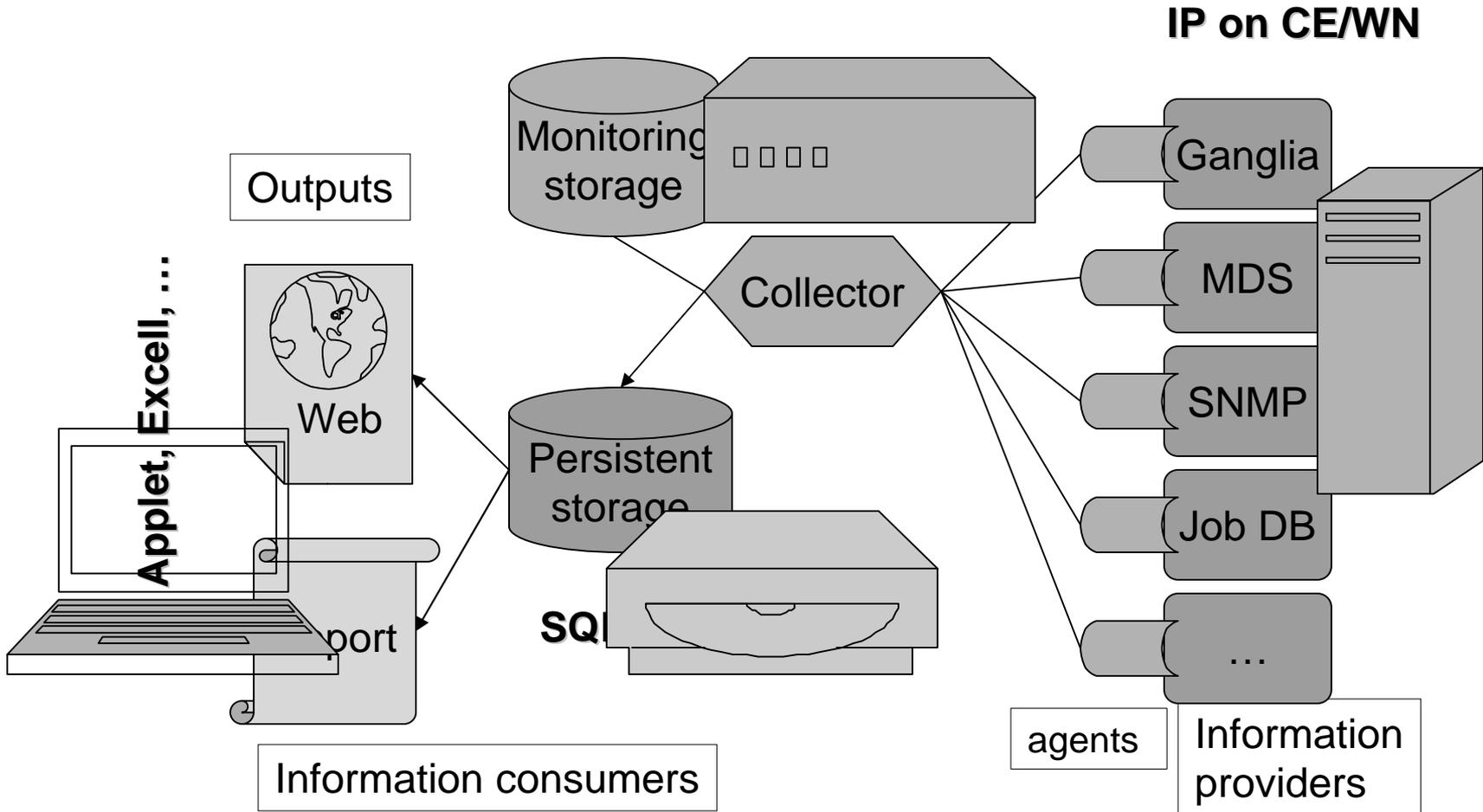
Outputs (for clients)



Monitoring framework



MF - Placement



Grid3 and LCG

- Grid3 should federate with the LCG
 - We will have a working group within Grid3 to understand what is needed for basic interoperability, specifically submission of jobs and movement of files across the grids.
- There will be other issues that may affect interoperability
 - consistent replica management, virtual organization support and optimizing of resource usage across federated grids.
 - We do not have the effort to address all these during the Grid3 project itself. We will identify, discuss and document such issues for collaborative work with the LCG.
- Many of the working areas in Grid3 are already joint projects between the LCG and Trillium or the S&C projects.
 - Additional collaboration in areas of monitoring and operations have been discussed over the past few months.
- As we proceed to better understand the technical plans the expectation is that we will propose further areas of joint work.
- **Use lessons learnt for DC2**

