EUAsiaGrid: A standard computational platform for e-Science

Ludek Matyska
CESNET and Masaryk University
Czech Republic

Bio-workshop

Singapore, May 4th , 2010

A little bit of a background information

- **What is a *grid*?**
  - **A platform for scientific collaboration**
  - **A scientific tool to help with data manipulation and processing**
  - **A computational platform**
- **And how it compares with a *cloud*?**
  - **A source for computing and storage capacity**
  - **Flexible, easy to access resource**
- **And a *cluster*?**
  - **A building block for both grids and clouds**

Grid as a collaborative platform

- **State before the Grid:**
  - **Scientists/teams have resources or access to them**
  - **Teams are working independently, they do not share their resources (no technology support)**
  - **Data sharing on a very primitive level, e.g. secure copy (scp) between teams**
- **State with a Grid:**
  - **Team's resources are connected**
  - **Sharing is easy, both for data (complex support like catalogues, databases, ability to work directly with data (if allowed by data owner) in "not-owned" repositories**
  - **Scientists could focus on their science and not on the technology behind it**

Data manipulation and processing

- Concept of *Storage Elements* (SE) where data are primary deposited
  - Different access methods supported (not only scp anymore)
  - Could have complex internal organization (databases)
  - Access control—data owners could specify in a great detail who can work with each element of the data
- Metadata and catalogues
  - To make high level organization of data possible even if huge data are stored and processed
- Access to the computing resources
  - Access control and access methods compatible with the methods and protocols used by the compute fabrics

Computational platform

- Primary batch processing oriented
  - User prepares a job and submits it to the grid together with requirements on the computing node(s)
    - All necessary information is part of the *input sandbox*
  - Grid components analyze the requirements and try to find the Computing Element (CE) where the processing could really be done
    - Taking usually into account also proximity to data to be processed
  - Job is sent to the CE and eventually runs there
    - Due to the asynchronicity of the grid distributed infrastructure the CE may behave differently from what is expected by the grid major components
  - When job is finished, user can retrieve the results
    - Either from the *output sandbox* or from SE

Computational platform—resources

- The grid resources are made from what individual users and teams “bring” with them
- Naturally, independent resource providers can also contribute
  - However, this is usually a part of some strategy like “we want to become grid resource providers for the following scientific communities”
- Each resource owner does have a full control over its resources
  - A kind of “voluntary contribution” from the point of view of the grid infrastructure
  - But not “voluntary” from the point of view of the owner (usually does have a target community it is willing or required serve)
- Power of “rich” users/teams
  - They have they own resources, so they decide who other can be allowed to use them
- Curse of “poor” users/teams
  - Dependent on the decision of the resource owners

Organization

- The grid infrastructure is not centrally *managed*, but it is centrally *coordinated*
  - High independence of individual sites, willing to work together (but not under some supreme managerial hierarchy)
- Users and resources are organized in *Virtual Organizations* (VO)
  - Each user must be a member of at least one VO to be able to use the grid infrastructure
  - Each resource declares which VOs is supports (i.e., users of these VOs can use the resources)
    - Each site defines how much resources are available for each supported VO and also eventual restrictions and priorities

Desktop grids

- A more voluntary processing platform
- As the name suggests, *desktop computers* considered primary resource
- Loosely connected, no central coordination
- Benefits:
  - Everybody with a desktop can contribute and become a part of such a grid
    - Potentially really huge processing power available
- Problems:
  - Much lower reliability of individual site
  - A risk of site “cheating” (not doing of what promised)
  - A security/privacy risk (are you sure the site does not make its own copy of your data and programs?)
  - A computing model must be adapted to this environment

Clusters

- More or less standard servers with Intel compatible CPUs (IA-32, x86_64, IA-64 architecture) in one or more racks
  - Sometimes nodes connected via high speed network like Infiniband
- Mostly Linux operating system
  - MS Windows can be used as well, but currently very little support for appropriate grid middleware
- A headnode
  - Accepts jobs for worker nodes
  - Sometimes available for some small interactive processing (like job preparation, programs compilation etc.)
- Worker nodes
  - All other nodes of a cluster
  - May not have a direct Internet connectivity
    - In such a case data and jobs sent exclusively through the headnode
  - Usually not for interactive use

Clouds

- A rather recent commercial platform
- (Large) Pool of virtualized servers
  - Users submits not jobs, but full virtual machines with jobs inside them
- Targets real-time requirements
  - Fast deployment of new virtual server
  - Can quickly react on user's changing requirements
- Standard clouds rather simple
  - Easy to use web interface
  - No collaboration support (standard Consumer – Provider model, ideal for commercial use)

European Grid Activities

SINCE 2001, BIOMEDICAL APPLICATIONS HAVE BEEN DEPLOYED ON DISTRIBUTED COMPUTING INFRASTRUCTURES

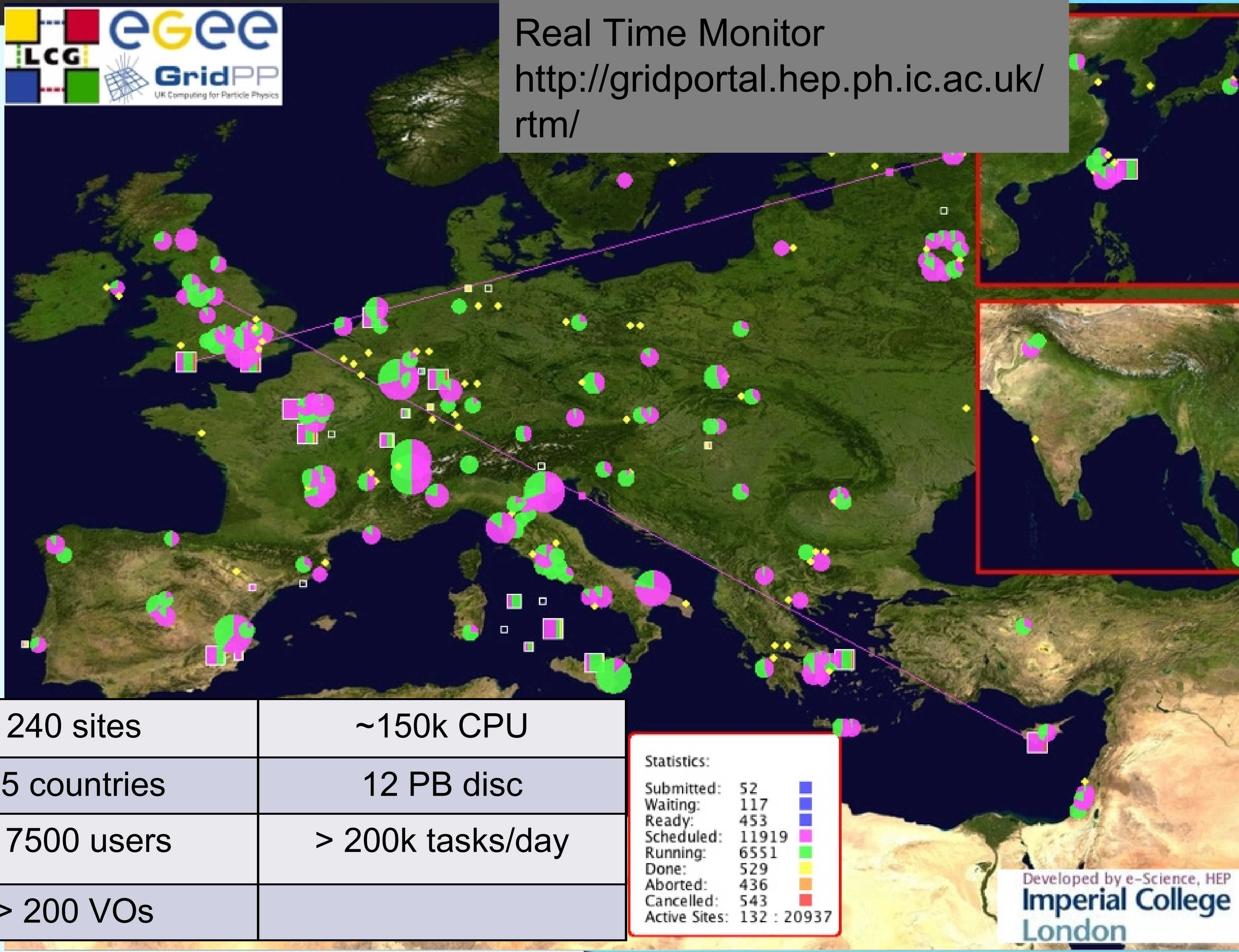
- **2001: DataGrid**
  - **Deployment of a prototype testbed**
- *2002: DEISA – **Distributed European Infrastructure for Supercomputing Applications***
- **2004: EGEE I –** Deployment of a European infrastructure for scientific production
- **2006: EGEE-II –** Further deployment of an infrastructure for scientific production

  ***Cloud Computing:** Amazon Web services (EC2)*
- **2008: EGEE-III:** Stable production infrastructure + migration to a federation of national grids

  DEISA2 – HPC Grid
- **2010- …: European Grid Initiative (EGI), project InSPIRE**

EGEE Grid Infrastructure

Real Time Monitor
http://gridportal.hep.ph.ic.ac.uk/rtm/



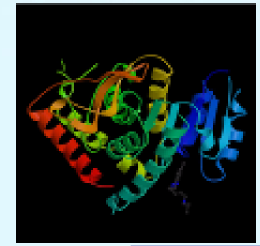
240 sites	~150k CPU
45 countries	12 PB disc
> 7500 users	> 200k tasks/day
> 200 VOs	

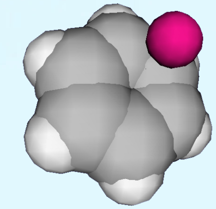
Application breakdown

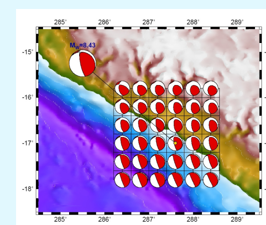
- **6 scientific areas are covered by EGEE grid**

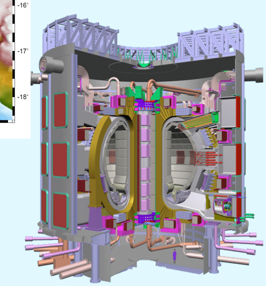
  < 100 applications deployed

	6/2006	2/2007	1/2008	2009
Astron. & Astrophysics	2	8	9	
Comp. Chemistry	6	27	21	
Earth Science	16	16	18	
Fusion	2	3	4	
High-Energy Physics	9	11	7	
Life Sciences	23	39	37	
Others	4	14	21	
Total	**62**	**118**	**117**	





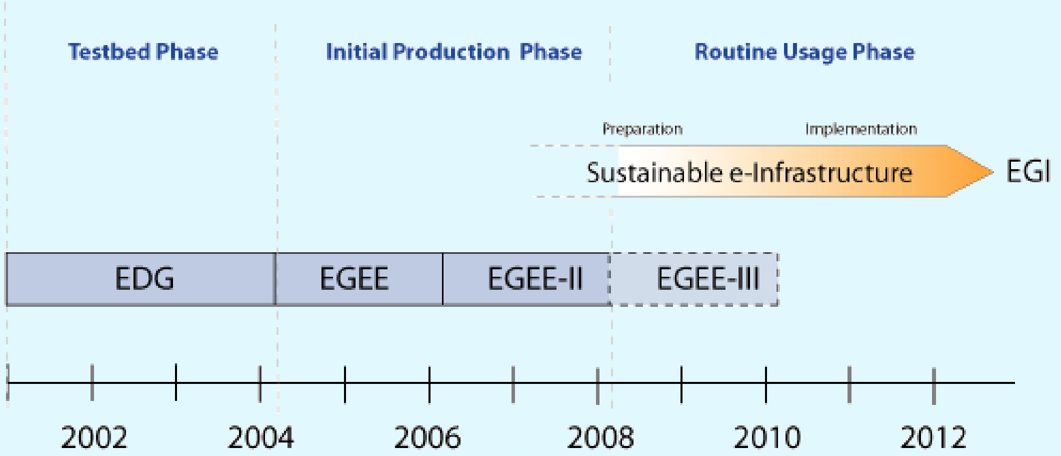




Condensed Matter Physics
Comp. Fluid Dynamics
Computer Science/Tools
Civil Protection

European Grid Initiative

1. **Goal:** long-term sustainability of grid infrastructures in Europe
2. **Approach:** establish a federated model bringing together National Grid Infrastructures (NGIs) to build the European Grid Infrastructure (EGI)
3. **EGI Organisation**: coordination and operation of a common multi-national, multi-disciplinary Grid infrastructure
   - ✓To enable and support international Grid-based collaboration
   - ✓To provide support and added value to NGIs
   - ✓To liaise with corresponding infrastructures outside Europe



EUAsiaGrid project and background

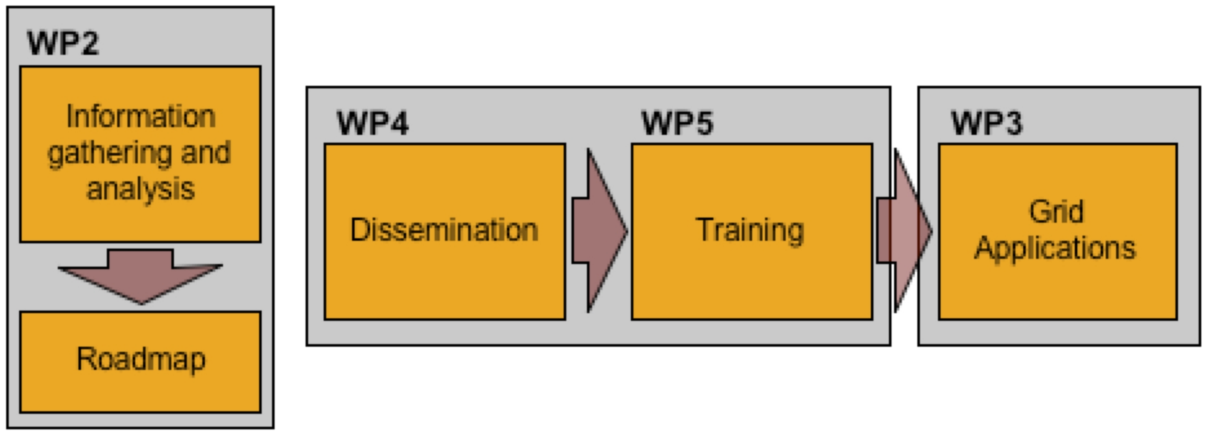
- An EU co-funded project to help create grid environment in the Asia-Pacific region
  - From April 1st 2008 to June 30th 2010
- Oriented towards provision of a “standard” grid, based on the gLite middleware (EGEE-like)
  - Open also to other grid platforms, but they are not its primary focus
- The goal is NOT to build the grid infrastructure, but to provide support for its use
  - Major effort goes to application support and training
  - Especially, the EUAsiaGrid project does not directly pay for any resources

Partners

1 *Istituto Nazionale di Fisica Nucleare (Italy) (coordinator)*
2 *CESNET (Czech Republic)*
3 *University of Manchester (United Kingdom)*
4 *HealthGrid (France)*
5 Ateneo de Manila University (Philippines)
6 Australia National University (Australia)
7 Academia Sinica (Taiwan)
8 Advanced Science and Technology Institute (Philippines)
9 Hydro and Agro Informatics Institute (Thailand)
10 Infocomm Development Authority (Singapore)
11 Institute of Information Technology (Vietnam)
12 Institute Teknologi Bandung (Indonesia)
13 National Electronics and Computing Technology Center (Thailand)
14 University Putra Malaysia (Malaysia)
15 MIMOS Berhad (Malaysia)
16 Institut de la Francophonie pour l'Informatique (Vietnam)
17 National University of Singapore (Singapore)

Breakdown in work packages

Work Package	Title	WP Coordinator
WP1	Project management	INFN (M. Paganoni)
WP2	Requirement capture and coordination policy definition	UNIMAN (A. Voss)
WP3	Support of scientific applications	CESNET (L. Matyska)
WP4	Dissemination	ASGC (V. Huang)
WP5	Training	INFN (M. Fargetta)



Project Challenges

- First EU supported grid project in the Asia-Pacific region
- First project to cover several countries at the same time
  - Geographically and culturally diverse area
  - No common framework like the European Commission in the EU
  - Very different previous grid experience
- Interest in desktop grids

Project Challenges – applications and users

- **Extensive application area coverage**
  - **"Classical" versus "New" or "Emerging" application areas**
    - **High Energy Physics, Bioinfomatics, Computational Chemistry versus Disaster Mitigation, Social Science, Cultural Heritage**
- **A cycle of *dissemination -> training -> support* (for production usage)**
- **Focus on international collaboration**
  - **Within the Asia-Pacific region and between AP and EU**
- **Create an environment where grids are used as regular production tools for science (and elsewhere)**

Sustainability

- Abused word in EU
  - Means to have guaranteed funding for the infrastructure at a higher level than just through independently submitted and evaluated projects
  - Requires governmental or similar level interest
- A Roadmap for the further development of grids in the Asia-Pacific region
  - Using EU experience in setting up EGI
  - However, creating National Grid Initiatives as basis for the international collaboration more difficult in the Asia-Pacific region
  - Lessons learned from the APAN (AP Networks infrastructure)

Project in the 1st year

- Setup of basic AP grid infrastructure
  - Mostly based on the fact that Academia Sinica in Taipei serves as the Tier 1 (highest level) site for the WLCG (HEP) infrastructure and has long term experience from EGEE series of projects
  - Several new sites setup and certified
- Creating a catch all VO EUAsia
  - Generic approach, not tailored for a specific scientific area
    - Support for wide range of applications
  - Users could join regardless on their scientific interest
  - Uses the basic AP grid infrastructure
    - Cross border collaboration simplified

Dissemination and training in the 1st year

EUAsiaGrid

- **Extensive dissemination activities**
  - **Project organized workshops (usually also with training activities)**
  - **iSGTW as the major media partners**
  - **Active participation on many scientific conferences, with focus on non-ICT/non-Grid ones**
  - **Journal publications**
- **Training focusing**
  - **Future trainers ("train the trainers" principle)**
  - **Users/scientists (including help with their problems)**
  - **A concept of "keep and eye" on trained scientists**

Project in the 2nd year

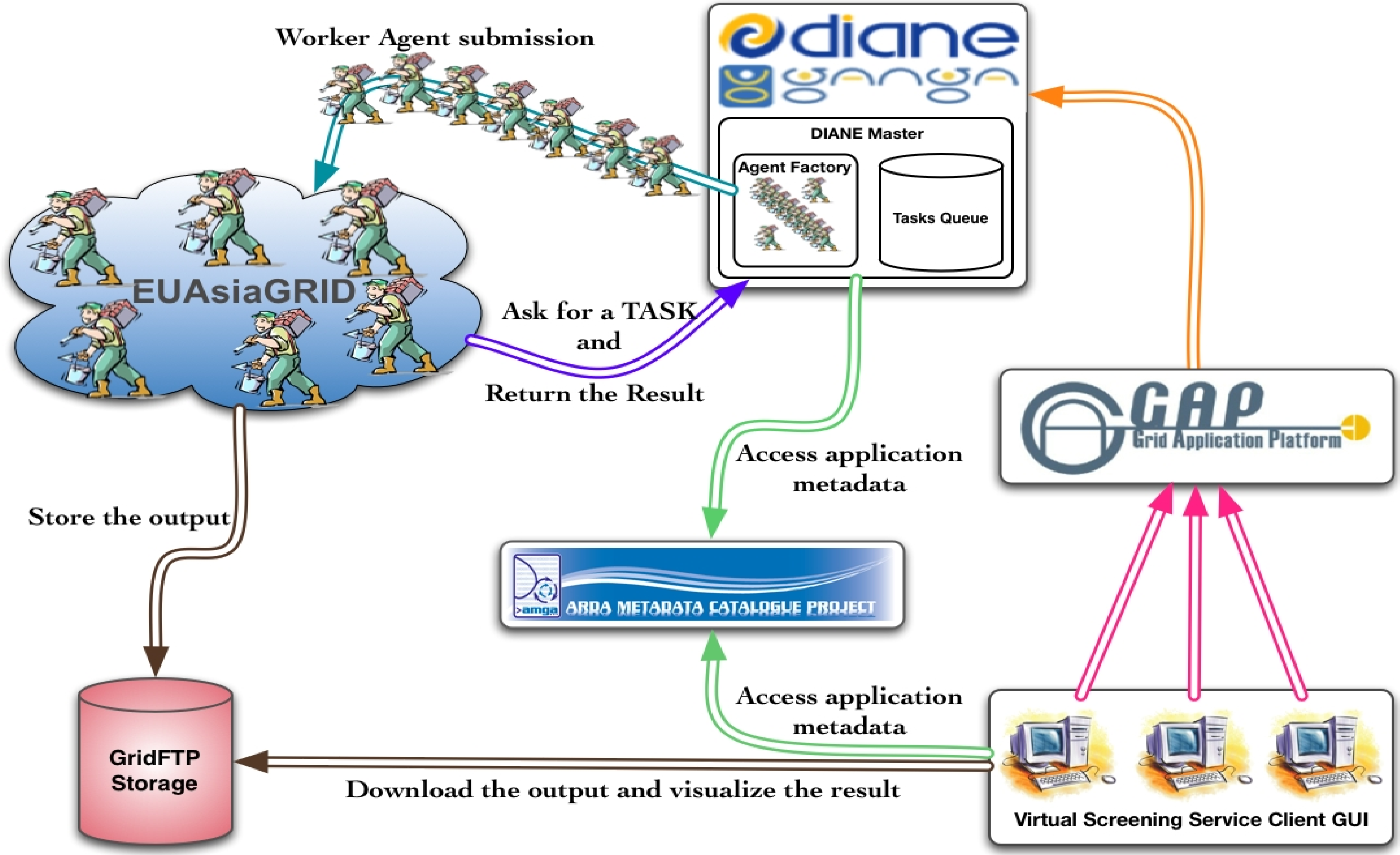
- Large data/compute experiments (challenges)
  - Avian Flu DC2 refinement
  - Dengue Fever drug discovery
  - Earthquake and climate change simulations
- Support for complex workflows
  - Needed for complex simulations
- Focus on new application domains
  - Social sciences, disaster mitigation
  - Does not mean other areas neglected
- More support for end users
  - Application repository
  - Application porting

EUAsia VO in details

- *Generic*, application neutral VO
- Starting from ASGC and UPM, currently also resources from EU and AP partners
- Some 600 CPU and 65 TB available
- Each partners/country established a User Interface to access the resources

Name of Institution	CPU(Cores)	Disk(GB)
IT_INFN-CT	100	460
CZ_CESNET	72	8754
ID_ITB	1	15
MY_UPM	126	983
MY_UM	24	134
MY_MIMOS	140	1246
PH_ASTI	56	10
PH_ADMU	1	30
TW_AS	100	51200
TH_HAII	2	1083
TH_NECTEC	4	1474
VN_IFI	8	733
VN_IOIT	4	733

EUAsiaGrid Virtual Screening Service



Earthquake simulation

- Objectives
  - Accurate simulation using dense network data plus historic data with a a precise simulation model
  - Seismic wave propagation analysis to reduce potential impacts
  - Planning for real time hazard mitigation
- Methodology
  - Implements and use seismic wave propagation model in the EUAsia VO grid environment
  - Validate results, implement case studies on disaster mitigation on a per country base for a start
  - Seismic network connection (TW, PH, VN, ID)

Earthquake Simulation – collaboration

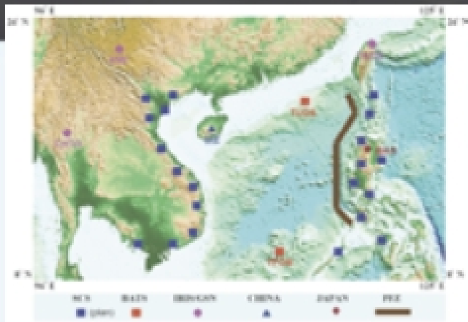
	Sensor Network	Seismic Wave Prop. Analysis	Seismic Data Center
Partners	VN, PH, ID, TW		
User Community	Philippine Institute of Volcanology and Seismology (PHIVOLCS), Vietnamese Academy of Science and Technology, The Incorporated Research Institutions for Seismology (IRIS), Global Seismic Network (GSN), Institute of Earth Science & National Central University, Taiwan, Local, Regional, and Global Disaster Mitigation Organization.		
Tech Maturity	TW- Most dense; VN- Sensor Stn ready; PH- expanding	Analysis Model and knowledge available; Cluster and gLite Resources in place; From Global model toward higher resolution regional/country model;	SeisGrid@TW, IRIS, GSN
Exemplar	Integrated Sensor Network by VN, PH and TW	Without local geological data, accurate analysis is not achievable. High resolution historical TW earthquake data sets.	Federation of available Data Centers

Earthquake simulation – plans

- Facilitate understanding of ground motion mechanism, rupture process, velocity structure and topographic characteristics
- GeoFramework as the basis
  - gLite plus GAP plus DIANE plus GANGA
  - Portal for data and analysis services
- Earthquake monitoring and simulation
  - Data Center prototype
  - Basic resource estimation:
    - cca 100 CPU per months
    - 10–30 TB
- Collaboration with Italy and Greece in the EU
  - Similar problems and experience

e-Science for Earthquake Disaster Mitigation

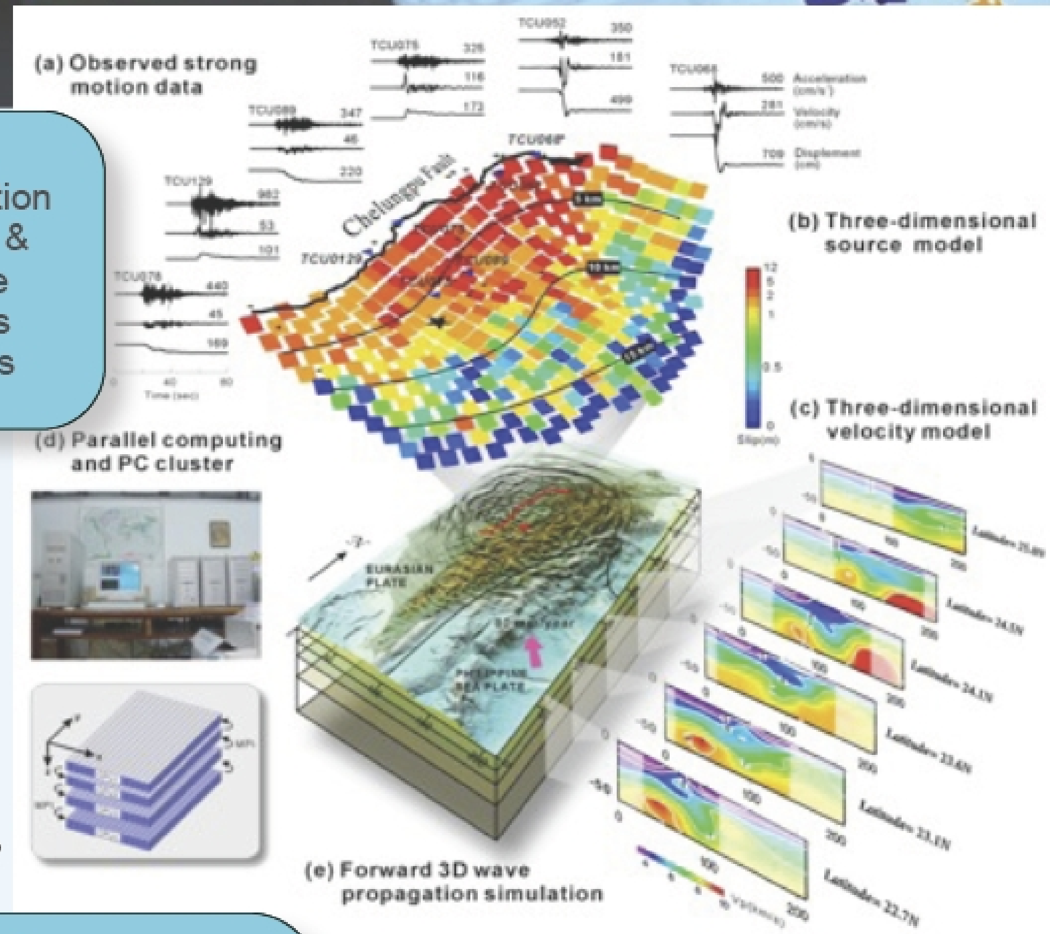
Seismic Sensor Networks



Local Sensor & Observation Data

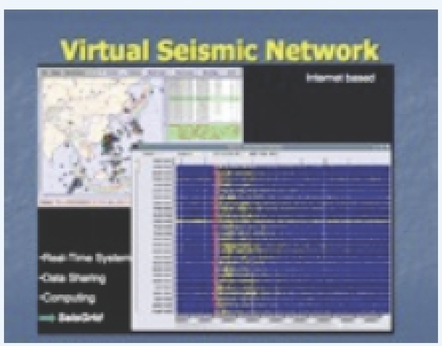
Global/Regional Sensor Data

High Resolution Source & Rupture Process Analysis



Fast Reporting System

Archive



Earthquake Data Center (SeisGrid)

Ref. Historical Events Data

Archive

Forward Simulation & Event Construction on Grid

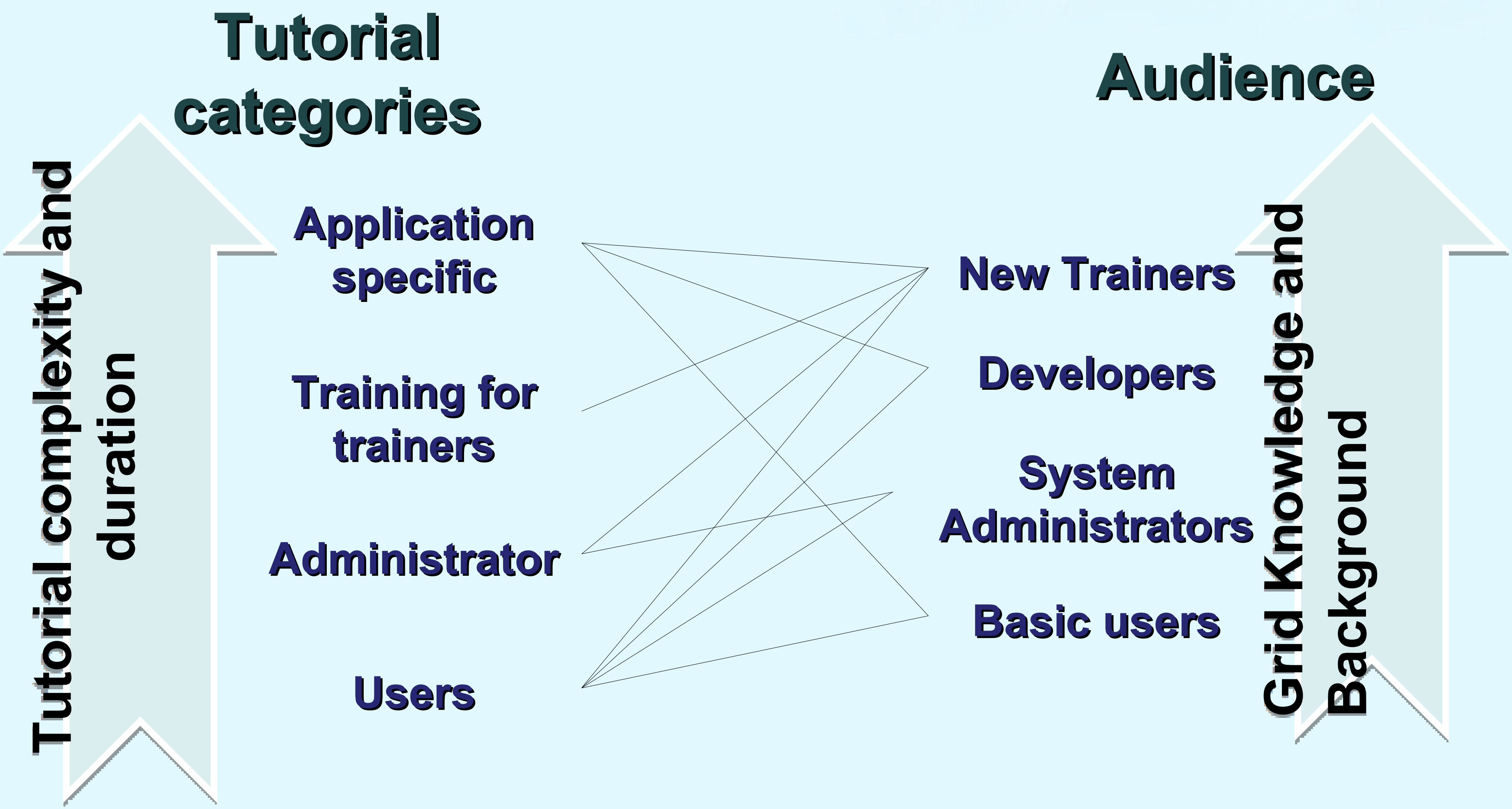
Risk Analysis & Reduction

FP7-INFRA-223791

7

FP7-INFRA-223791

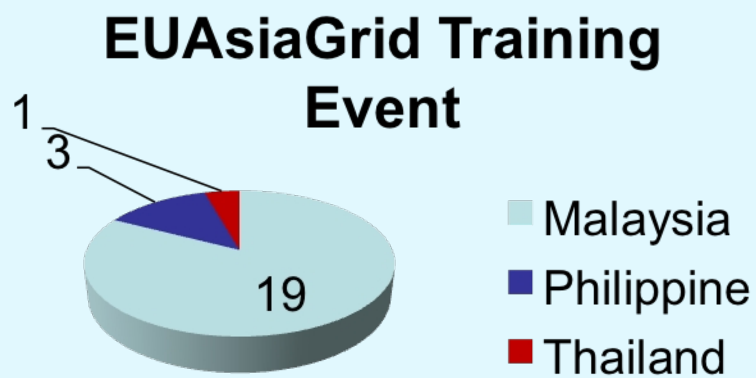
Training tailored to audience

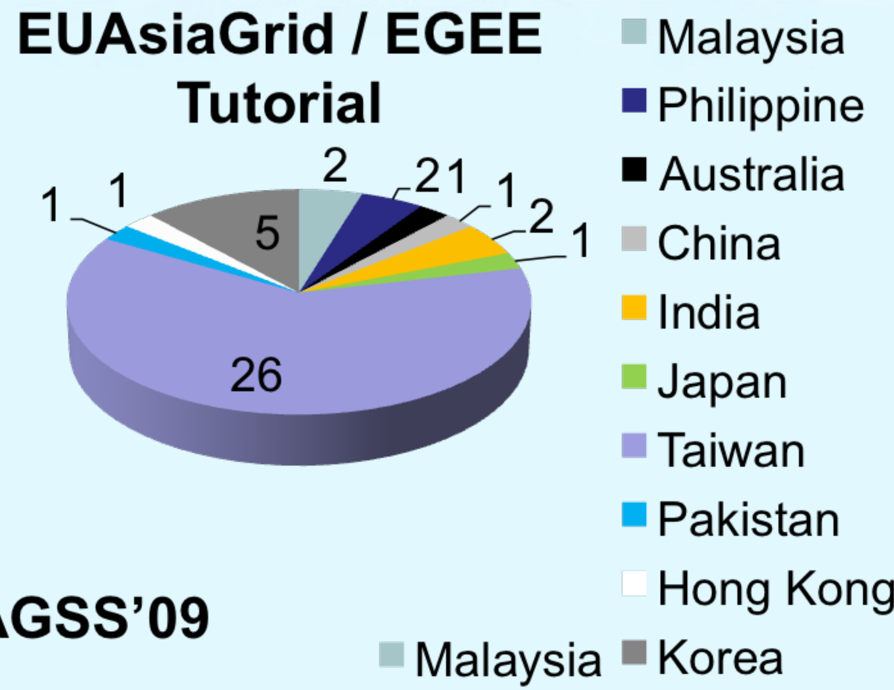


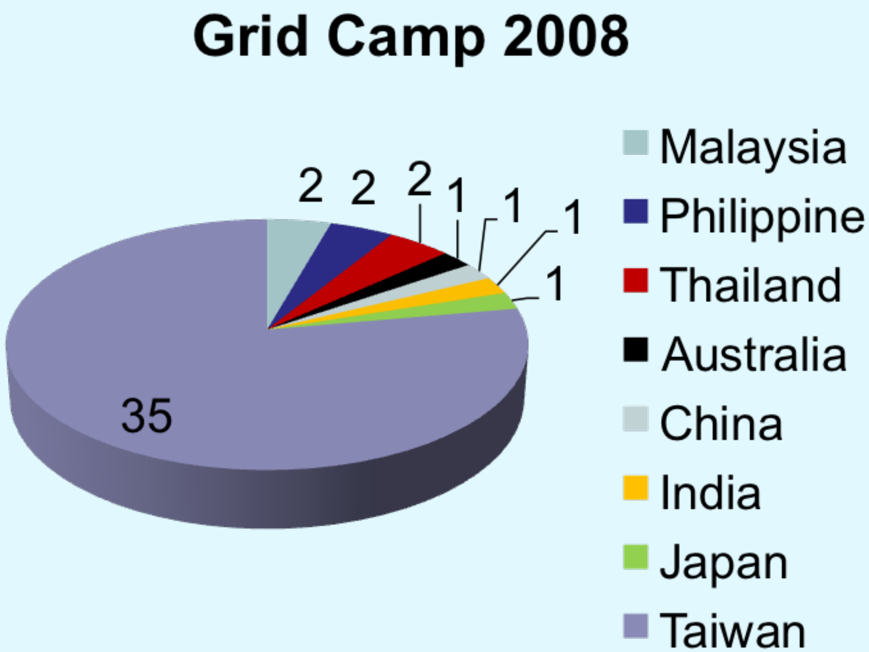
Training events

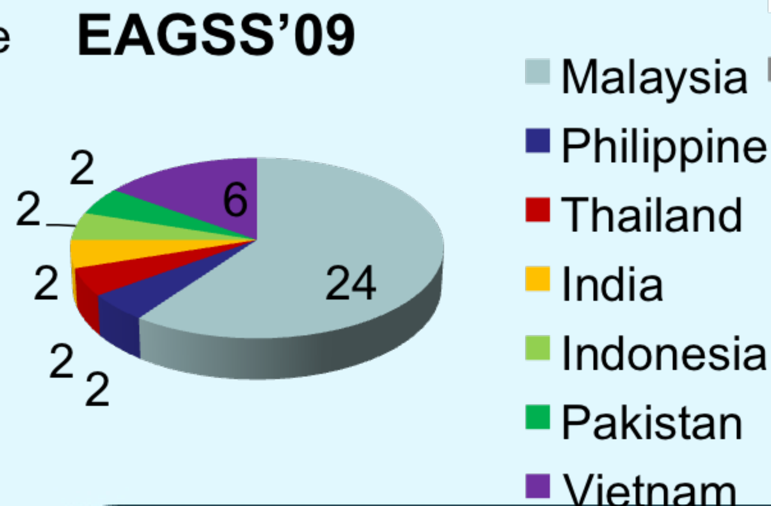
- **EUAsiaGrid Training Event at MIMOS**
  Kuala Lumpur, Malaysia, 28-31 Jul 2008
- **Grid Camp 2008 at ASGC**
  Taiwan, Taipei: 18-24 Oct 2008
- **EUAsiaGrid / EGEE tutorial at ASGC**
  Taiwan, Taipei: 18 Apr 2009
- **EUAsiaGrid Summer School**
  Kuala Lumpur, Malaysia, 27 Jul - 8 Aug 2009
- **EUAGSS09 (with CNRS): sysadmins + new grid users**
  Hanoi, Vietnam, 27 Sep - 03 Oct 2009
- **Train the trainers at ASGC**
  Taipei, Taiwan: 05-09 October 2009
- **ACGRID II (with CNRS): applications + train the trainers**
  Kuala Lumpur, Malaysia, 02-14 Nov 2009
- **gLite training for trainers at ASGC**
  Taipei, Taiwan, 5-7 March 2010
- **Social Simulation Tutorial**
  Taipei, Taiwan, 7 March 2010

Participation to training events









A vision for an Asia-Pacific e-Infrastructure for Research and Education

EUAsiaGrid

Roadmap

e-infrastructure

Roadmap in more details

- Provides first insight on how to organize EGI-like grid infrastructure in the Asia-Pacific region
- It is targeting
  - Policy makers and Funding agencies
  - Resource owners
  - Scientific communities (their self-organization)
- Input from similar activities in Latin America and South East Europe
- Looks for benefits of public-private partnership in the Asia-Pacific region
- Interested partners of EUAsiaGrid created consortia to participate in the EU CHAIN project
  - It will provide next step, following the situation in a truly global approach (CHAIN has partners around the globe)

Conclusion

- EUAsiaGrid project laid down foundations for the grid infrastructure in Asia-Pacific region
  - The organizational and funding aspects covered in the Roadmap
  - The infrastructure as a resource for the EUAsia VO
  - The training activities with a multiplicative effect (Training the trainers)
  - Dissemination material (tangibles, papers, presentations)
- Grid infrastructure and organization compatible with the EU EGI activities
- Continuation of efforts within the CHAIN project

Thanks for your attention!

Ludek.Matyska@cesnet.cz

http://www.euasiagrid.eu/