Managing Enterprise Applications in Grid

Anirban Chakrabarti

Software Engineering and Technology Labs (SETLabs)

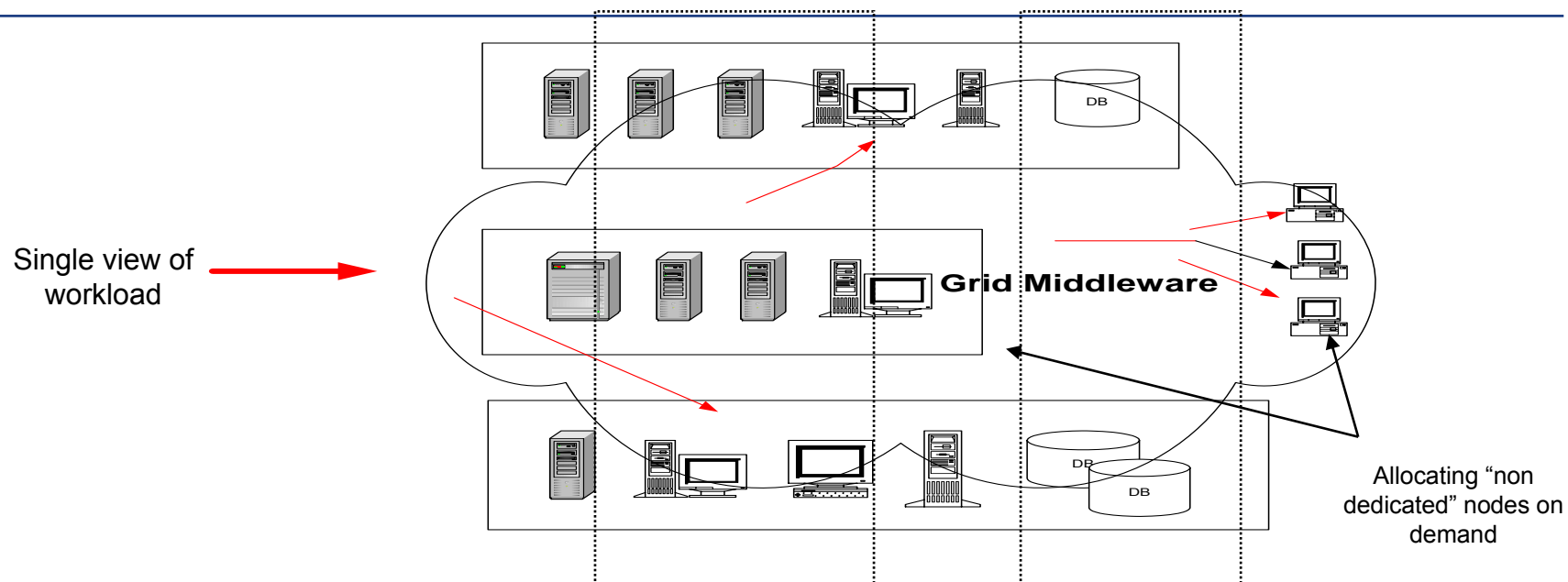Infosys Technologies

Infosys®

Agenda

- Grid Computing

- Enterprise Grid Computing

    – Issues

- Grid Computing at Infosys

    – Prototypes and Case Studies

Grid Computing Definition

We define Grid *as a software and hardware infrastructure, that allows flexible and seamless sharing of heterogeneous network of resources for compute and data intensive tasks and provides faster throughput at lower costs*

Infosys®

Improve your odds with Infosys Predictability

Grid enabled infrastructure – sharing of resources

Single view of workload

DB

Grid Middleware

DB

DB

Allocating "non dedicated" nodes on demand

- Infrastructure and data sharing through dynamic creation of virtual resource pools

- Non-dedicated (shared) nodes "on demand"

- Larger scope for infrastructure optimization

- Better efficiency and overall throughput can be achieved.

Benefits of Grid Computing

- Increased Resource Utilization

- Distributed Computing over heterogeneous resources on demand

- Virtualization of Resources – infrastructure and data

- Provisioning, co-allocation, accounting and metering of resources

- Resource Balancing

- Better Throughput

- Better Return-on-Investment

Infosys®

Improve your odds with Infosys Predictability

Typical Applications of Grid Computing

Industry	Applications
Automobile/Aerospace	Clash simulations, modeling
Life Sciences	Drug discovery, Gene matching, BLAST, Indexed search
Financial Services	Risk Management, Credit Analysis, Portfolio Analysis, analytic job processing
Geo Sciences	Reservoir modeling, Seismic analysis, Simulations
Rendering	Computer animation, Image processing,
Academia	SETI@Home, Astronomy

Main Concerns in Grid Computing

Concerns

Application Engineering & Management

Legacy applications need to be gridized. Gridization & Management frameworks are required
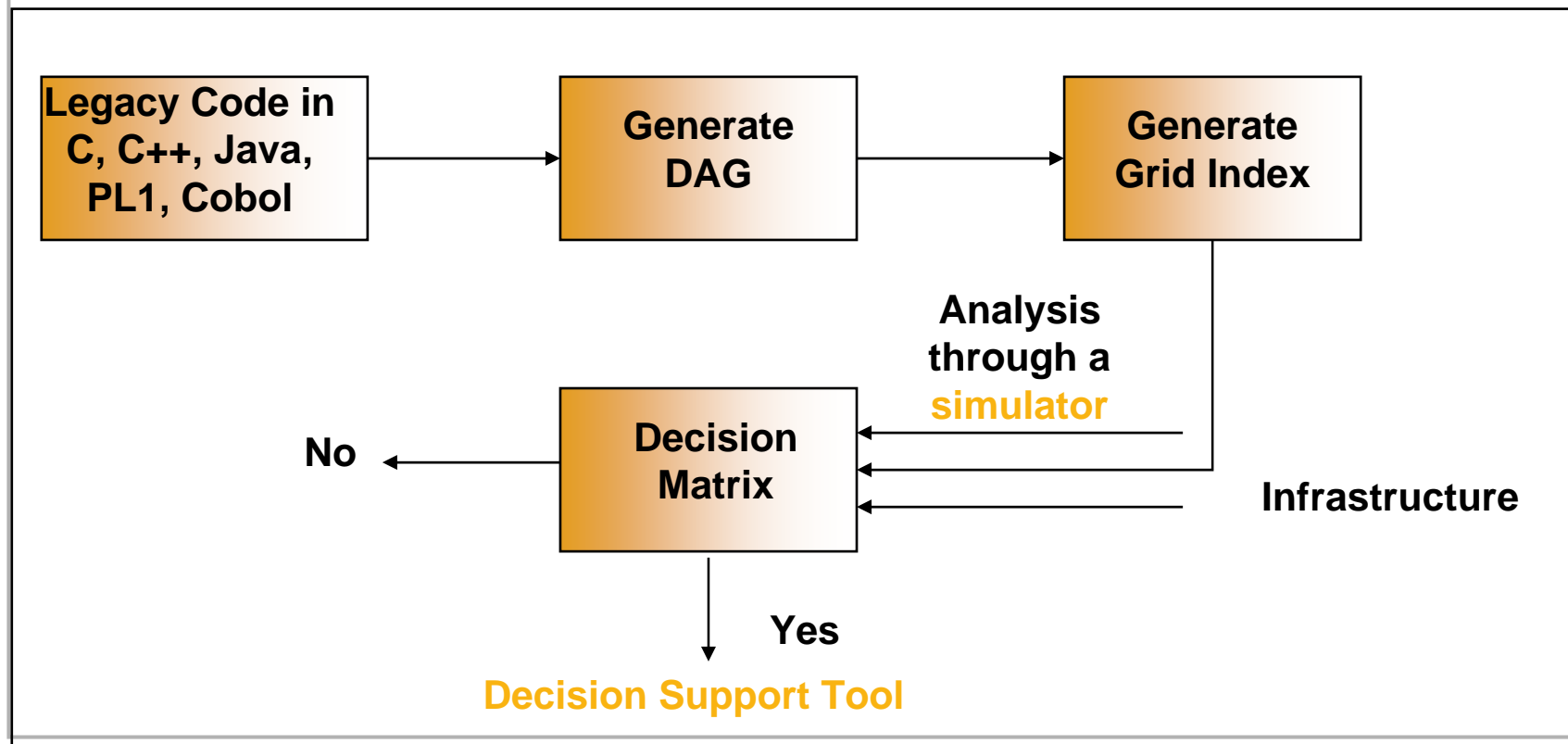
Security & Licensing

Issues regarding infrastructure, host and policies need to be taken care of

Application Engineering – Importance

- There is a tradeoff between computation and communication
  - All applications do not perform equally well, when gridized
  - Development of tool which quantifies the Grid Index
    - Computation/Communication tradeoff
    - Development of Grid index
    - Performance need to be quantified
    - Need for a simulator
- There is a need for categorization of applications
  - Need to develop body of knowledge
  - Templates & Best practices

Application Analysis

Templates   **Best Practices**   **Body of Knowledge**

Application Categorization

Legacy Code in C, C++, Java, PL1, Cobol → **Generate DAG** → **Generate Grid Index**

Analysis through a simulator

No ← **Decision Matrix**

Infrastructure

Yes

Decision Support Tool

Grid Management

- Managing Diverse systems
  - Common reporting
  - Autonomic Capabilities
- Managing SLAs
  - Guaranteed services needed
- Automating jobs and applications

Pharma desires innovation

- The launch of a new drug takes a minimum of 12-18 years
- R&D spend per new chemical entity approved is around $600 - $800MM
- Pharma desires domain and technological innovation to improve efficiency process and enhance productivity

RAPIDLY RISING DRUG R&D SPENDING IS HELPING USHER IN A NEW R&D PARADIGM
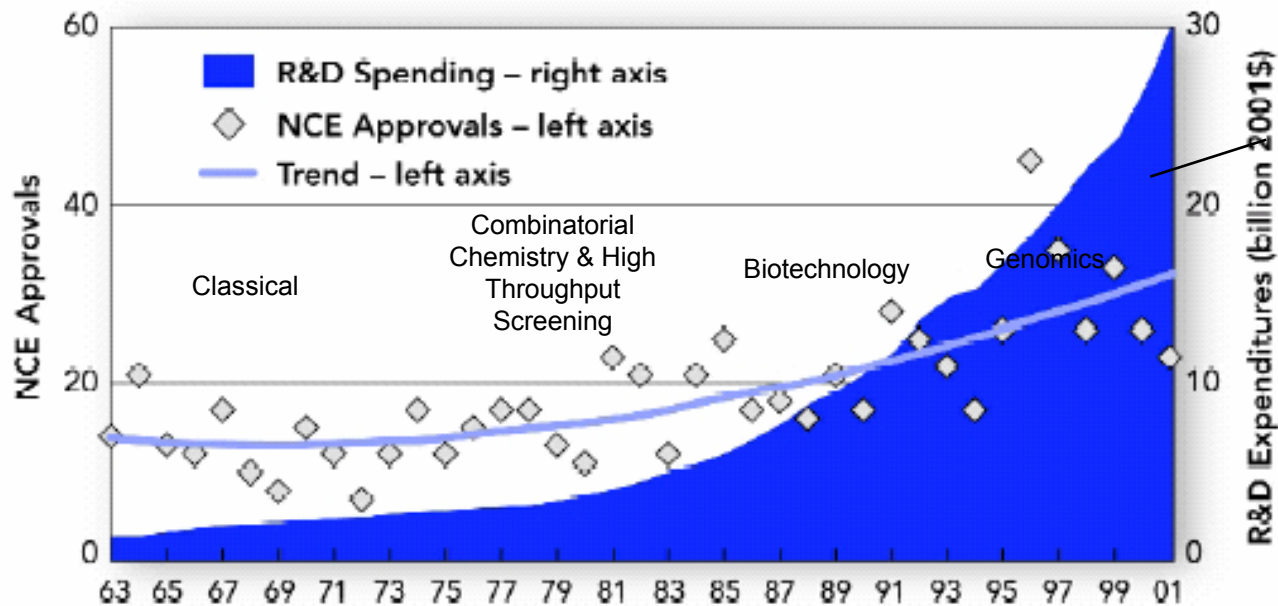
New drug approval have doubled in the past decade, the annual R&D spending has increased 12 times

Source: Tufts Center for the Study of Drug Development, PhRMA

Automation Scope in Pharma Industries

Desirable features which will streamline a workflow

- Integrate pipeline of tasks which involve algorithms, batch scripts, vendor applications, web services etc.
- Automate the manual data processing
- Integrate data to create a virtual repository
- Workflows need to be executed in bulk
- Assimilate outputs into a report
- Common access to reports
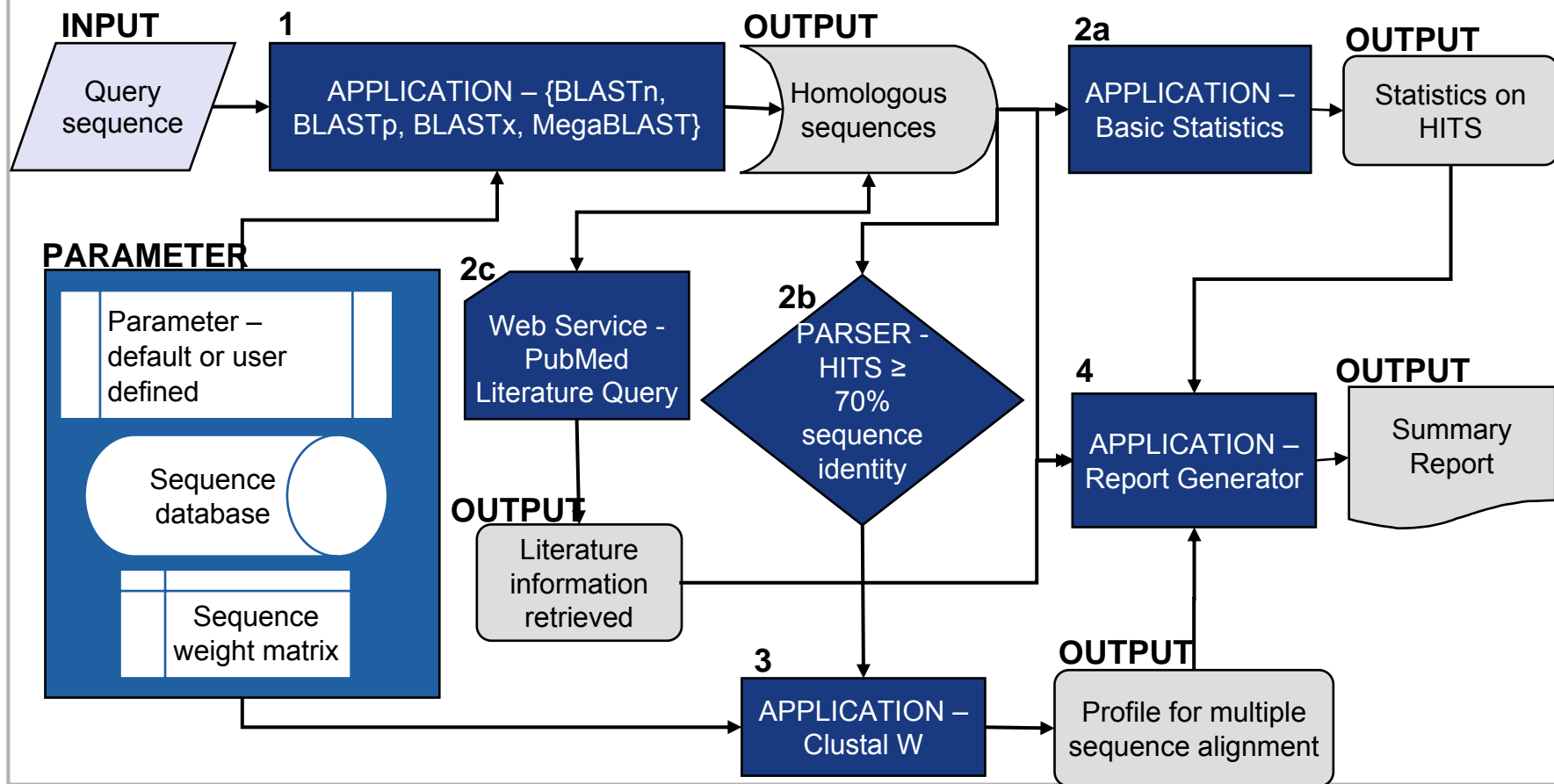- Secure exchange of data
- Audit trail of tasks

Which processes in drug discovery there is scope of creating a workflow which will impact productivity in research?

- – Gene annotation labs
- – Micro-array labs
- – Structural biology labs
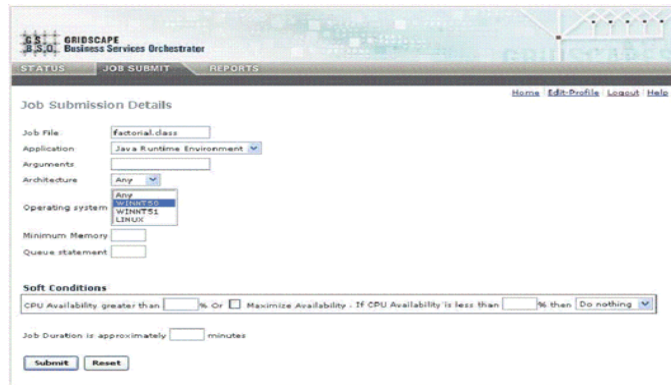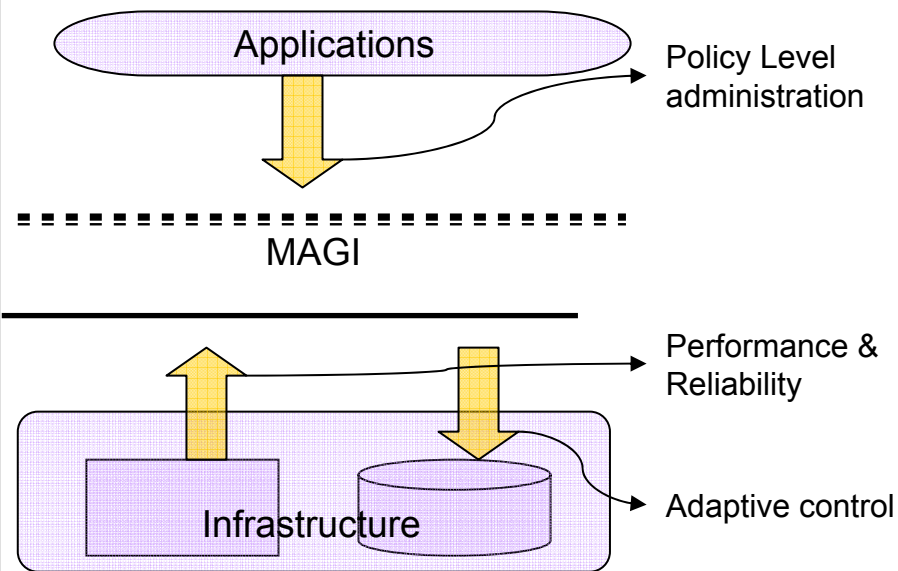- – High throughput screening labs
- – Toxicology labs
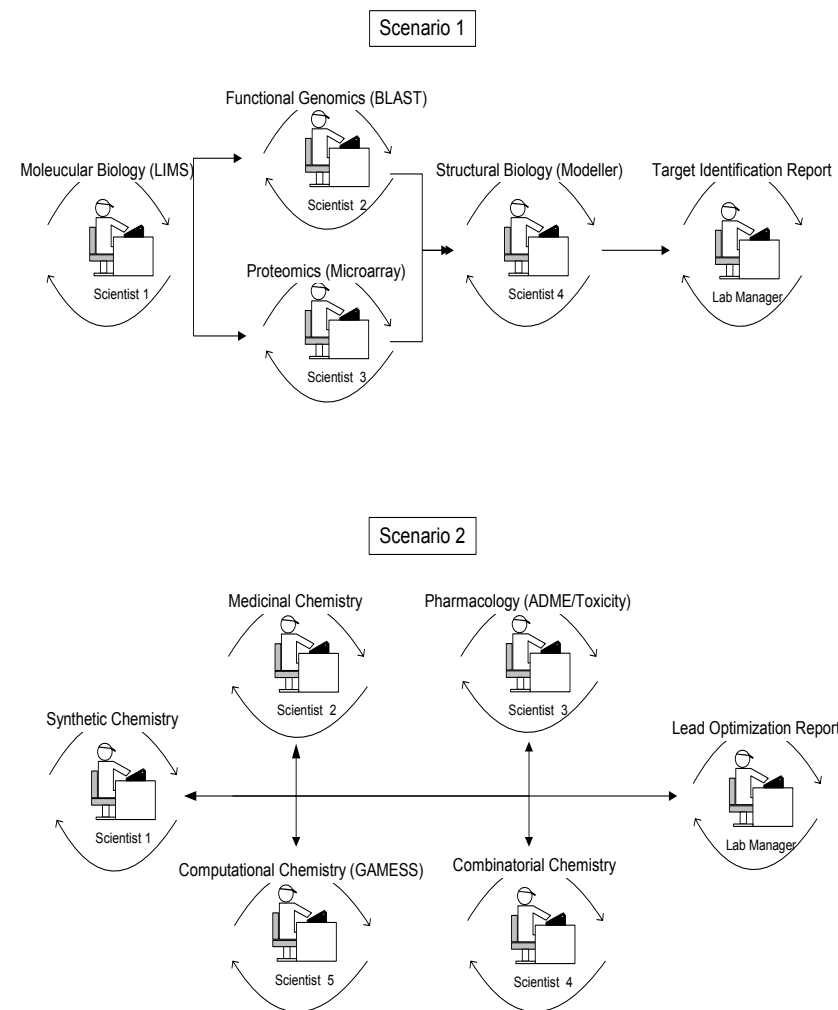
Design for the gene identification research workflow

Discovery Research workflow – gene annotation program includes BLAST, CLUSTAL W or HMMER algorithms, PubMed web services, batch scripting, statistics, conditional looping, reports.

MAGI- Management of Adaptive Grid Infrastructure*

Applications

Policy Level administration

MAGI

Performance & Reliability

Infrastructure

Adaptive control

* Patent pending

Overview

MAGI provides an integrated approach towards resource management through virtualization, dynamic policy management and autonomic capabilities –collectively known as business services monitoring (BSM)
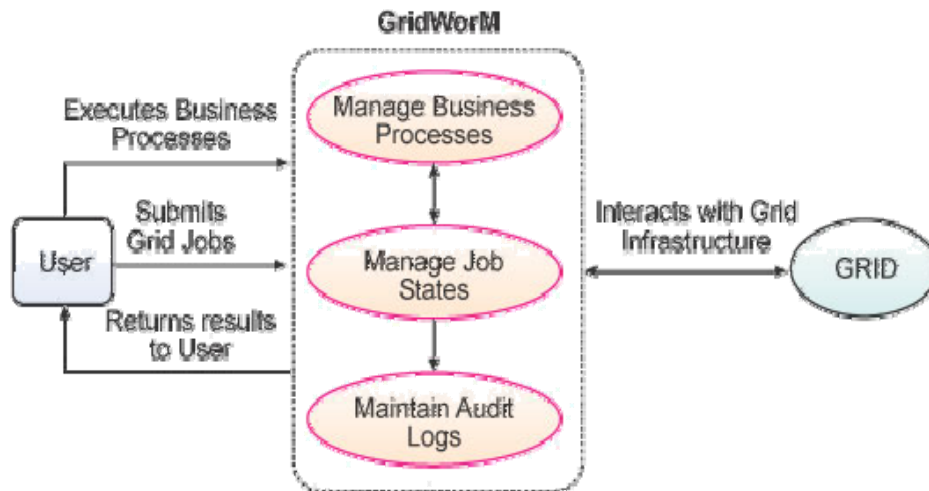
Motivation

- Improved scheduling of batch jobs based on data collected from hardware infrastructure
- Autonomic features to minimize user intervention
- Ease the burden of the administrators

Use Cases

- Business level management of grid systems
- managing change in IT infrastructure
- managing performance

Grid Workflow Manager*

GridWorM

Description

GridScape Workflow Manger (**GridWorM**) is a generic workflow manager to execute workflows with relationships among processes/jobs through a workflow
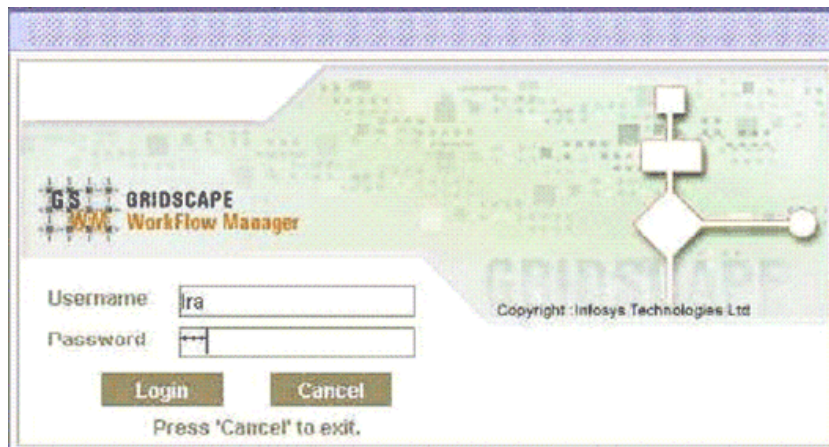
Motivation

To increase the automation of processes

- Several industries require an automated process of job submission to improve the productivity

To reduce the turn around time of process flow

- Need to efficiently distribute processes/jobs to the underlying infrastructure
- Need to efficiently leverage the existing data parallelism of processes/jobs

* Patent pending

GAEF - Application Engineering Framework for Grid

Grid Application Engineering Framework is a set of tools that analyzes the application portfolio, profiles the code base for Grid enablement, and suggests best possible way of Grid enabling the code
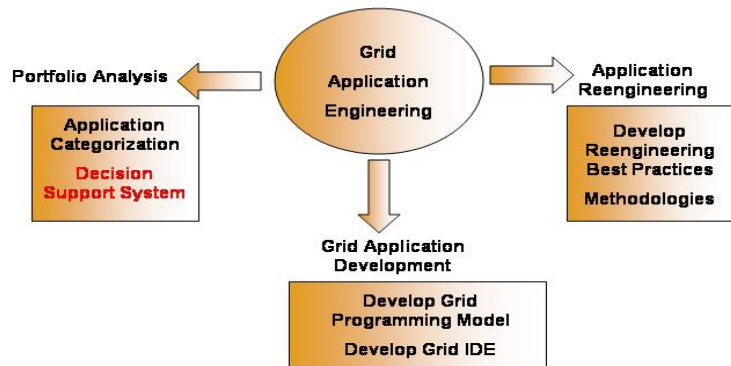
Features

- Source code analyzer (both static and dynamic analysis)
- Profiling code to identify components for parallelization.
- Grid simulator to simulate parallel execution of code in target Grid environment to analyze the benefits.
- Prescriptive methodology to support code migration to the Grid environment.

Grid App Engg Framework
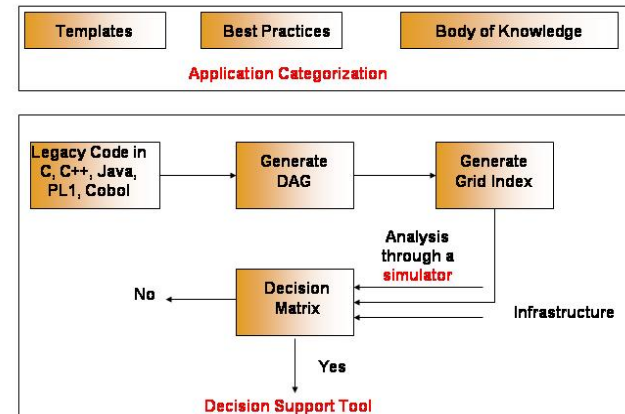
Portfolio Analysis

Grid enabling J2EE Applications

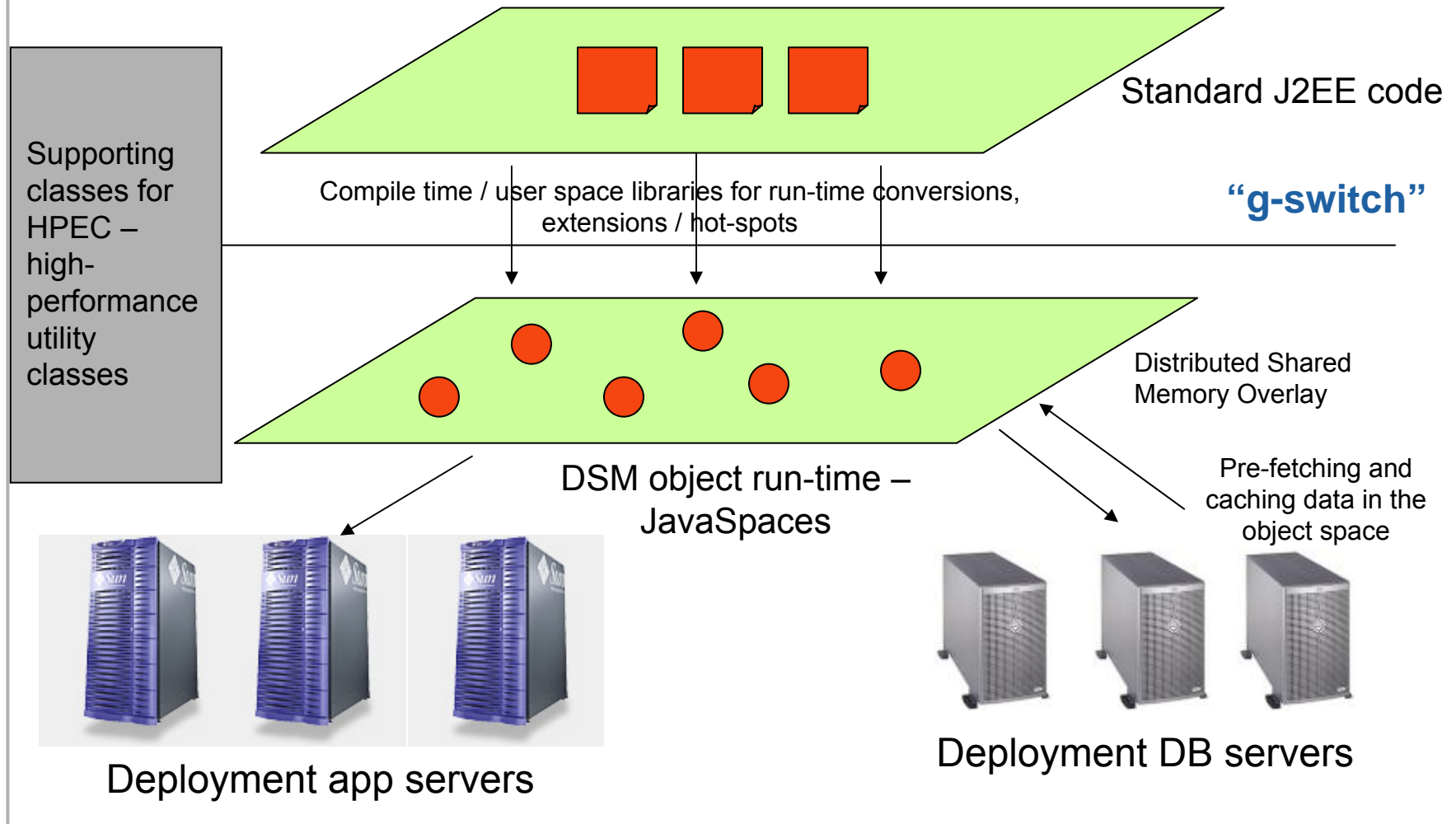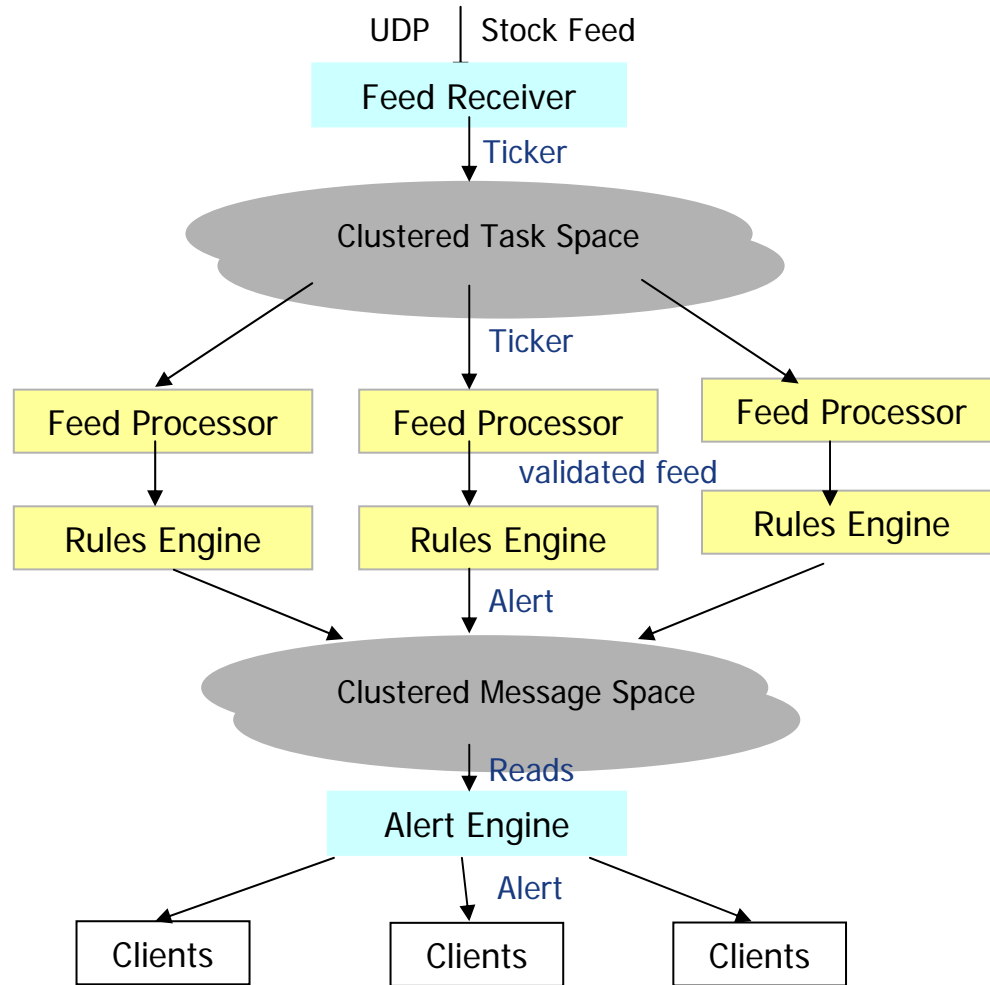- The J2EE programming paradigm has widespread deployment in enterprise computing in mission critical applications. The paradigm is prevalent in both on-line transaction processing (OLTP) applications, batch mode enterprise backend applications and compute / data intensive business logic programs (analytics).
- The heightened needs of performance, scalability, availability overwhelms today's application architecture:
  - Today's high volume transaction sites often have to cater to millions of transactions per second with sub-second response time
  - Self services intranet / internet portals deals with thousands of live transactions with reconciliations at a highly scalable and available manner.
  - The back-end processing and analytics load like fraud detection, credit check, risk management, pricing have also increased exponentially

- Clearly, the traditional approach of achieving high-performance, reliability, scalability and availability is often brittle and comes at a very high cost of provisioning.
- Grid based fabric can be used as a backbone for getting extreme scalability and performance

HPEC - Grid enabling J2EE applications

Standard J2EE code

Supporting classes for HPEC – high-performance utility classes

Compile time / user space libraries for run-time conversions, extensions / hot-spots

"g-switch"

DSM object run-time – JavaSpaces

Distributed Shared Memory Overlay

Pre-fetching and caching data in the object space

Deployment app servers

Deployment DB servers

HPEC - JavaSpaces based Market Data Processor

UDP | Stock Feed

Feed Receiver

Ticker

Clustered Task Space

Ticker

Feed Processor   Feed Processor   Feed Processor

validated feed

Rules Engine   Rules Engine   Rules Engine

Alert

Clustered Message Space

Reads

Alert Engine

Alert

Clients   Clients   Clients

Motivation

- Real time alerting
- High volume & high speed market data

Features

- Load balancing using Spaces based architecture
- Implemented with Javaspaces and Jini based Gigaspaces EAG
- Failover for feed processors

GridOS – Bringing Virtualization to the Grid

GridOS is a light weight Operating System that will bring the benefits of Virtualization to the Grid environment. Grid jobs will be run on virtual machine instances instead of on the physical server.
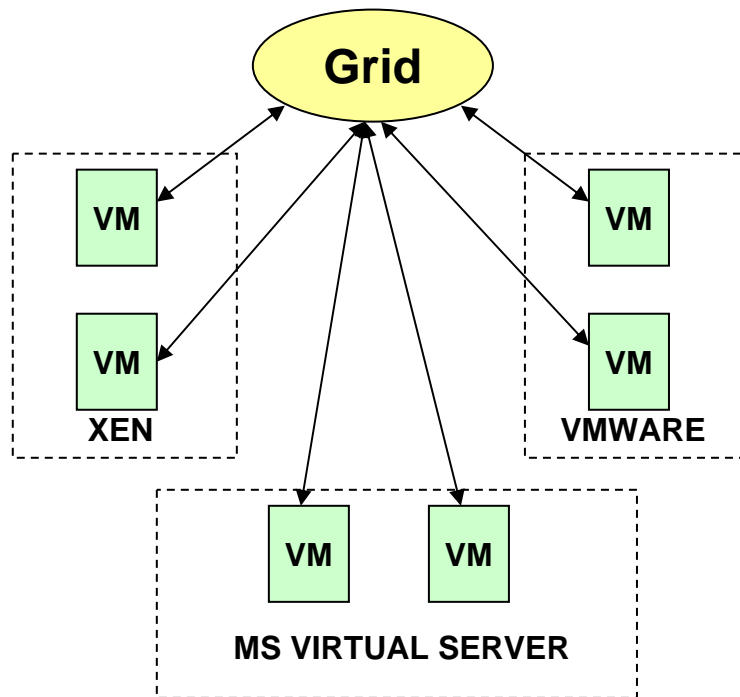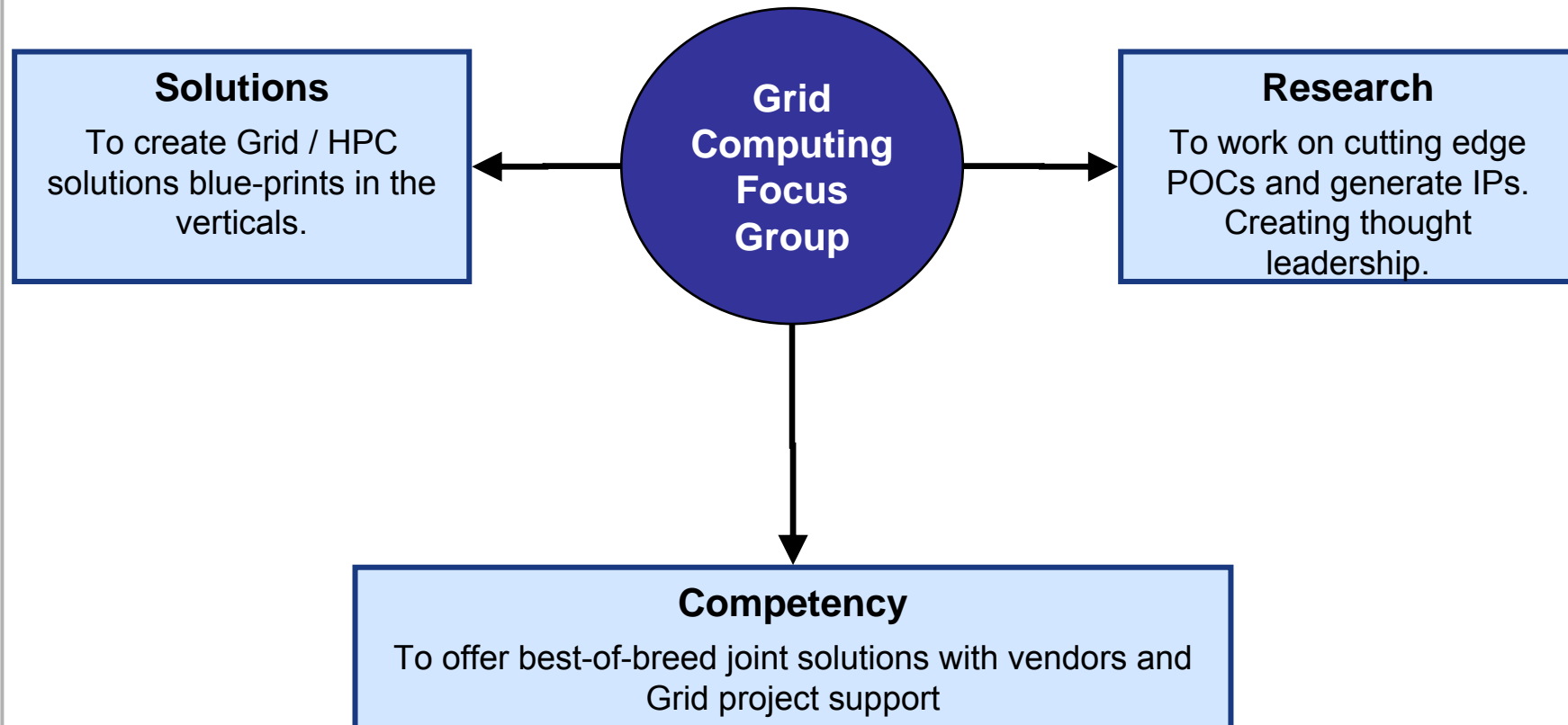
Features

- Efficient provisioning of resources
- Effective isolation between VM instances
- Minimal performance impact due to virtualization
- Also ensure minimal initialization overhead
- Fault tolerance and job migration
- Metering and accounting
- By providing accurate metrics on resource usage

Benefits

- Aids in Server Consolidation thus increasing utilization and reducing total cost of ownership
- Protects Grid applications on compute node from malicious applications and isolates Grid jobs through sandboxing
- Provides isolation, resource guarantees and security to Grid jobs

Grid

VM
VM
XEN

VM
VM
VMWARE

VM
VM
MS VIRTUAL SERVER

* VM – Virtual Machine

SETLabs Grid Computing charter

Solutions

To create Grid / HPC solutions blue-prints in the verticals.

Grid Computing Focus Group

Research

To work on cutting edge POCs and generate IPs. Creating thought leadership.

Competency

To offer best-of-breed joint solutions with vendors and Grid project support

Improve your odds with Infosys Predictability

Grid Computing Research Areas at Infosys

	Middleware & Workflow	Licensing & Security	Application Engineering	Virtual Data Management	Fabric Virtualization
Themes	*Middleware & Workflow*	*Licensing & Security*	*Application Engineering*	*Virtual Data Management*	*Fabric Virtualization*
	Policy Level Management	**Metering & Accounting**	**Grid Services Framework**	**Data Replication Strategies**	**Server Consolidation**
	Autonomic Capabilities	**Security Policy Management**	**Application Migration Framework**	**Dynamic data integration**	**Virtual Execution Engine**
Research Initiatives	**Grid Workflow**	**License Management**		**Meta data management**	**Resource Management**
	Advanced Scheduler	**Cross Domain Integration**			**Fault Tolerance**

Thank You

Anirban_chakrabarti@infosys.com