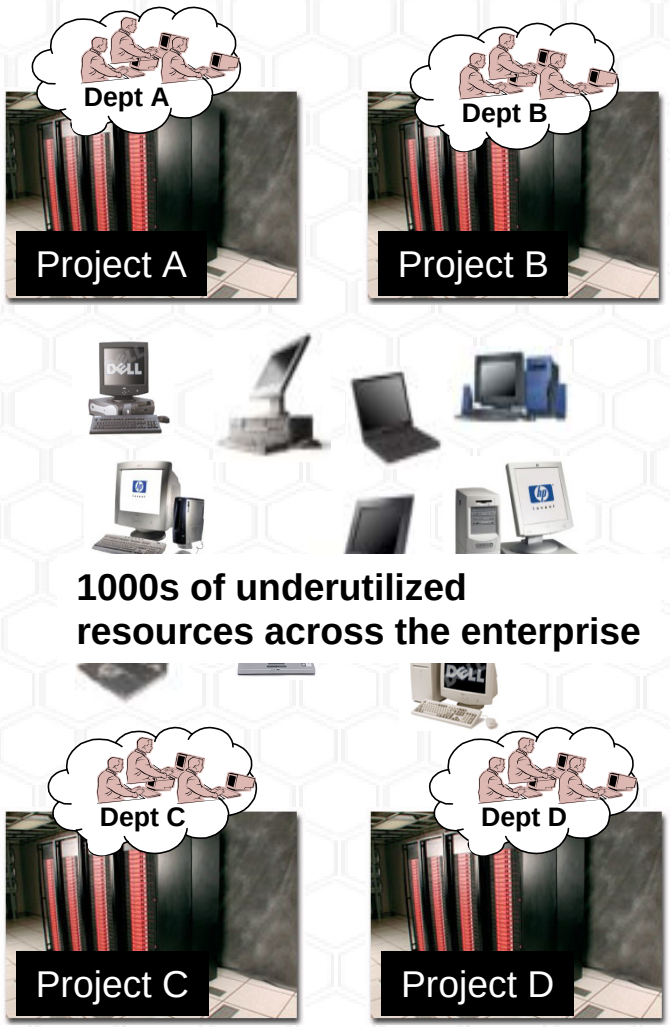
Grid Computing in the Enterprise with the UD MetaProcessor

Jikku Venkat
Chief Technology Officer
United Devices, Inc.
http://www.ud.com/
05 September 2002

Outline

- The High Performance Computing landscape today
- A compelling alternative opportunity – Enterprise Grid
- Some of the key challenges to overcome
- The MetaProcessor Architecture
- Solution Case Studies
- The future and a final summary

High-Performance Computing Today



- Departmental usage in industry; more sharing in universities
- Low utilization and duplication of compute and data resources
- Expensive resources: Supercomputers, SMP, Linux cluster
- Are the applications appropriate?
- Obsolete in 2-3 years

Massive amount of under-utilized capacity – mostly PCs

Relative processing capacity

	 1989 Cray Y-MP8/4128	 Intel® Pentium™ 4 Processor
CPUs	4 166MHz	1 P4 1.5GHz
Max RAM	1 GB	1 GB+
MFLOPS	~1333	~1800
Cost	~ $14,000,000	Under $300
	3/20/89 Electronic News	11/00 www.hardwarecentral.com

… and interconnect bandwidth, storage

1000s of underutilized resources across the enterprise

- All PC's are now connected
  - Bandwidth at 10, 100 and rapidly evolving to 1000 Mb/sec
- PC hard drive sizes at 20+ GB
  - Large amounts of disk space go unused
  - Nearly 50% unused – per MS study

**Data from Microsoft SIGMetrics Paper ('98-'00)*

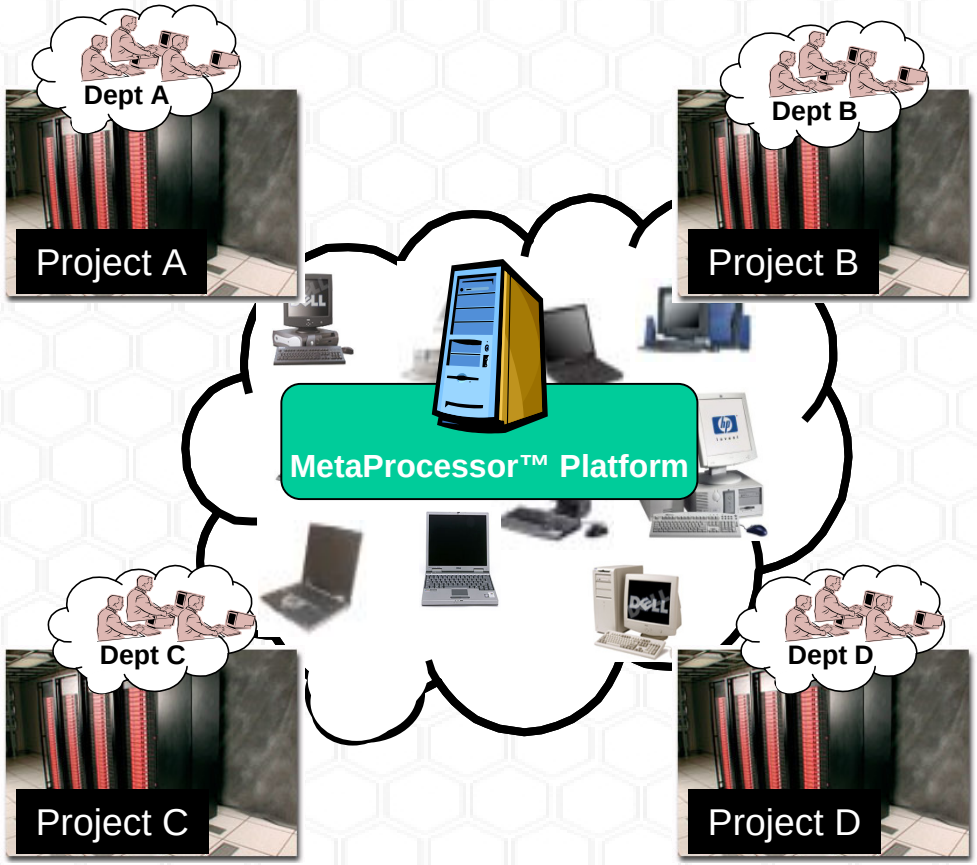
Availability, Utilization

1000s of underutilized resources across the enterprise

- 50% of all PCs available nearly 95% of the time
- 9% of PCs utilized nearly 100%
- Remaining 91% at or near idle
  - Median load is at 1-2%

**Data from Microsoft SIGMetrics Paper ('98-'00)*

Harness it! The MetaProcessor Extending the Grid



- 10x to 70x reduction in cost – millions of dollars in savings
- Nearly 100% resource utilization
- Enterprise-wide use for appropriate applications
- Preserve departmental control of policies and access controls
- Single virtual system view for application developers, administrators, users
- Asset appreciates over time – Moore's Law

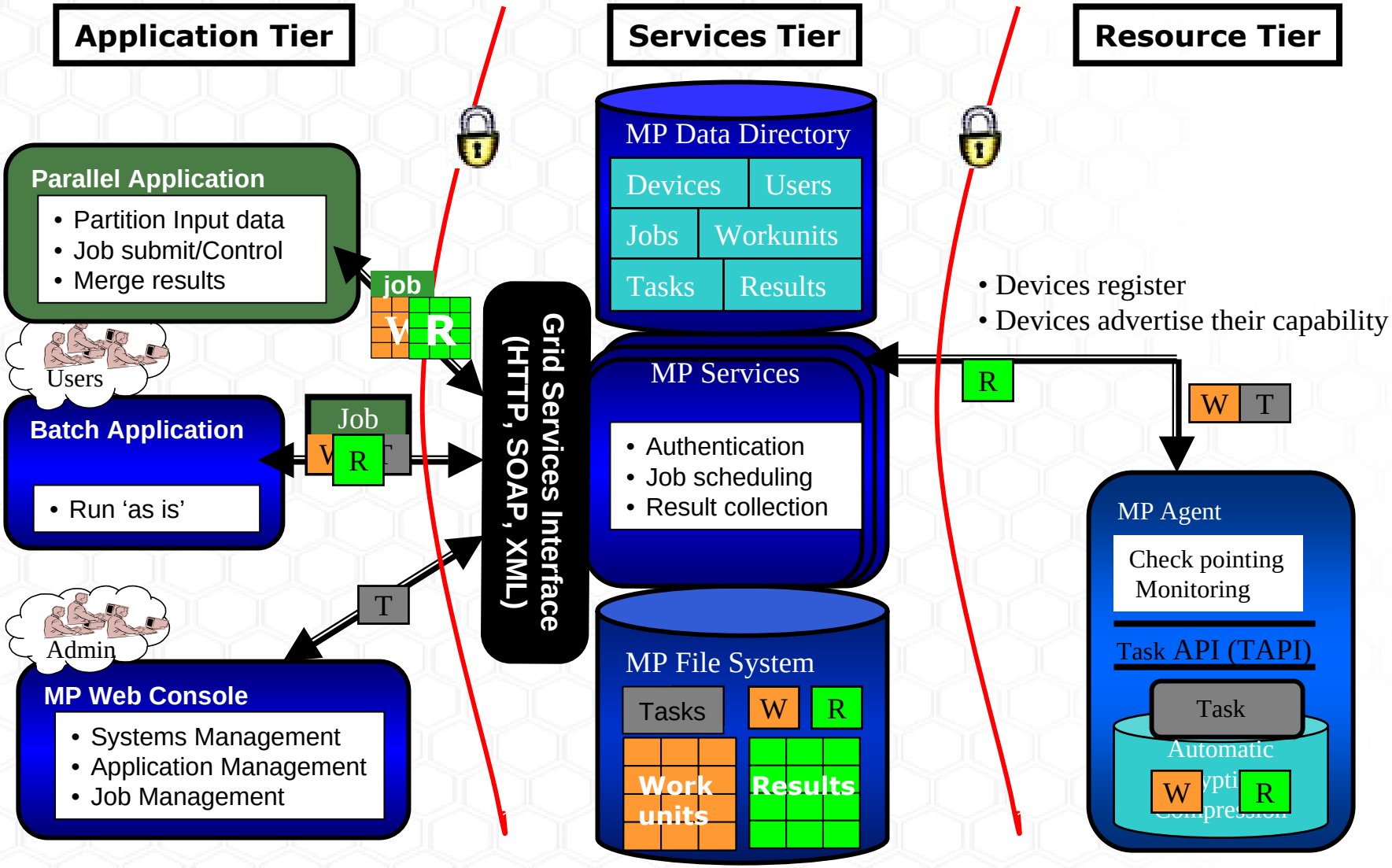
The Challenges

- **Security**
  - Providing controlled access to resources, applications, and data
  - Preserving the integrity of applications and data
  - Securing applications
  - Securing the application execution environment
- **Unobtrusiveness**
  - Will this be unobtrusive to the primary user of the PC?
  - How will this impact other users of the network infrastructure?
- **Scalability**
  - Will the resource scale to tens of thousands of devices? How?
  - What about application scalability?
- **Manageability**
  - How will this virtual system be administered and maintained?
  - How are reliability and availability addressed?
- **Dealing with non-dedicated resources**
  - Resources are not always available, connected, reliable

The Challenges

- **Application Feasibility**
  - This is feasible only for a small subset of embarrassingly parallel applications
  - How do we determine application feasibility?
- **Application Migration**
  - What is required to migrate applications to this grid?
  - Are source code modifications necessary?
  - How long does it take to migrate applications? What is involved?
- **Application Management**
  - How are applications updated as new versions become available
- **Data Management**
  - How are application datasets managed in the system?
- **Resource and Workload management**
  - Controlling and balancing workloads across a very large resource set
- **Application user interface**
  - Are application users aware of the grid?

MetaProcessor Architecture



Enterprise Grid / Public Grid / Partner Grid

Enterprise - MP

METAPROCESSOR™ PLATFORM

Aggregated Resources

Firewall

Public Grid
UD GMP

METAPROCESSOR™ PLATFORM

Firewall

Aggregated Resources

Public Internet Resources

Enterprise Resources

Partner Grid

METAPROCESSOR™ PLATFORM

Firewall

Internet For Coms Only

Firewall

Aggregated Resources

Security

- **Authentication**
  - User access to MetaProcessor requires an identifier and password
  - SSL-like protocol for authentication and network encryption key generation
  - Unique device identifiers and network session keys for authenticating devices
- **Encryption and Checksum**
  - Network communications encrypted using triple-DES network key
  - All data stored on devices is encrypted using triple-DES device key
  - All files stored on the devices are tamper-proofed using checksum validation
- **Organizations and Roles**
  - Users and applications allocated into organizations
  - User roles based on four levels of access control to the system
- **Digital Signature**
  - Application executable modules may be signed and validated on devices
  - DSA signature keys can be modified for each customer
- **Constrained execution environment**
  - MP Agent executes tasks in a sandbox with limited access to device resources

Unobtrusiveness

- **Preference profiles**
  - Control computation and communication time windows
  - Control disk space usage on devices
  - Control tasks that can execute on devices
- **Agent deployment**
  - Installation does not require a machine reboot
  - Deployed silently using enterprise software distribution tools
  - Run as a protected process such as, ‘WinNT Service’ or a user level application
- **Optimal usage of resources**
  - MP agent has a negligible memory foot-print
  - Lowest priority supported by the operating system
  - Limit disk usage based on device preference settings
  - Optionally configured to run in ‘screen saver only’ mode
- **Network unobtrusiveness**
  - Network communication only after task completion
  - All data transmitted on the network is automatically compressed
- **Optional User control**
  - User can snooze the task on non-dedicated devices
  - User can shutdown MP agent

Product Scalability

Devices	Connections per hour, Data-In/hour	MP Dispatch Service	Management Service	Application Service	MP Database
500	**83 conn/hr, 4 MB/hr**	**1 Server – 1CPU, 1GB RAM, 50 GB disk, Linux**			
1,000	**167 conn/hr 8 MB/hr**	**1 Server – 1 CPU, 1GB RAM, 20 GB disk, Linux**			**1 Server-1cpu, 1GB RAM, 50 GB disk, Linux**
10,000	**1,667 conn/hr 83 MB/hr**	**1 Server-1cpu, 1 GB RAM, Linux**	**1 Server-1 cpu, 512 MB RAM, 30 GB disk (shared w/UD servers), Linux**		**1 Server-2cpu, 4 GB RAM, 100 GB disk, Linux**
25,000	**4,167 conn/hr 208 MB/hr**	**1 Server- 2 cpu, 1 GB RAM, Linux**	**1 Server-1cpu 512 MB RAM 100 GB shared disk, Linux**	**1 Server-1cpu 1 GB RAM, 20 GB disk, Linux**	**1 Server-2cpu 8 GB RAM, 150 GB disk. Linux + Warehousing**
150,000	**25,000 conn/hr 1,250 MB/hr**	**3 Servers-2 cpu, 1 GB RAM, Linux**	**1 Server-2 cpu 512 MB RAM, 500 GB shared disk, Linux**	**1 Server-1cpu, 1GB RAM, 20 GB disk, Linux**	**1 Server–4cpu, 8 GB RAM, 150 GB disk, AIX + Failover + Warehousing**

back

Manageability

- **Web-based administration Console**
  - Remote administration of users, devices, applications and jobs
  - Features accessible based on role and access control policies
  - System management functions to control core services
- **Self-updating infrastructure**
  - Automatic update of MP Agent and tasks
  - Phases and versioning enable smooth transition to new code base
- **Automated systems management**
  - Periodic review and cleanup of stale data in the database and file system
  - Manager process restarts failed slave service processes
- **Rapid installation and upgrades**
  - Single command installation of MetaProcessor services in most environments
  - MP Agent compatible with most software distribution tools
  - Average time for MetaProcessor deployment is less than a day
  - Customized migration scripts to enable smooth upgrades

Dealing with non-dedicated resources

- **Optional redundancy parameters**
  - Each unit of work can be scheduled more than once
  - System can be set to wait for a minimum number of results to assure result integrity
- **Optional application-level checkpointing**
  - Provides features to enable applications to checkpoint their state periodically
  - System automatically resumes application execution from last saved state

Application Feasibility

- **Running many, many instances of the same application**
- **Coarse-grain data parallelism**
  - Data can be partitioned into independent chunks for executions
  - Large computation to communication ratios
  - Static data dependencies can be resolved
- **Coarse-grain control parallelism, pipelined execution**
  - The application itself can be partitioned into independent pieces
  - Output of one stage feeds the input of the next stage
- **Number of applications with this criteria**
  - Virtual screening of small molecules to develop drug leads
  - Genetic sequence analysis algorithms
  - Monte Carlo simulations
  - Decryption, key-breaking applications
  - Text mining applications
- **Windows executable exists**
  - Tools available to enable rapid migration – cygwin is an example

Application Management

- **Registration**
  - Programs are registered as *tasks* in the system
  - Registered tasks may be shared by one or more applications
  - Tasks have platform-specific executables called *task modules*
- **Versioning**
  - Tasks and task modules have versions
  - Versions enable tasks to be centrally upgraded and managed
- **Phases**
  - Tasks can be released in *test*, *pilot* and *production* phases
  - Phases enable developers & administrators to unobtrusively release tasks

Application Migration

- **MetaProcessor Task Wrapper**
  - packages executable programs without source code modifications. Runs ‘as is’
  - Transparent encryption and compression without source code modifications
  - Monitors and controls programs when executing on device

- **MetaProcessor Grid Services Interface (MGSI)**
  - Programmatic Web services interface based on HTTP, SOAP and XML
  - Support for 22 different programming languages
  - Minimal effort to develop *Application services*, tools and utilities

- **MetaProcessor Application Services**
  - Application services pre processes data, submit jobs and post processes results
  - Uses MGSI to interface with the MetaProcessor. Optionally can use batch utilities

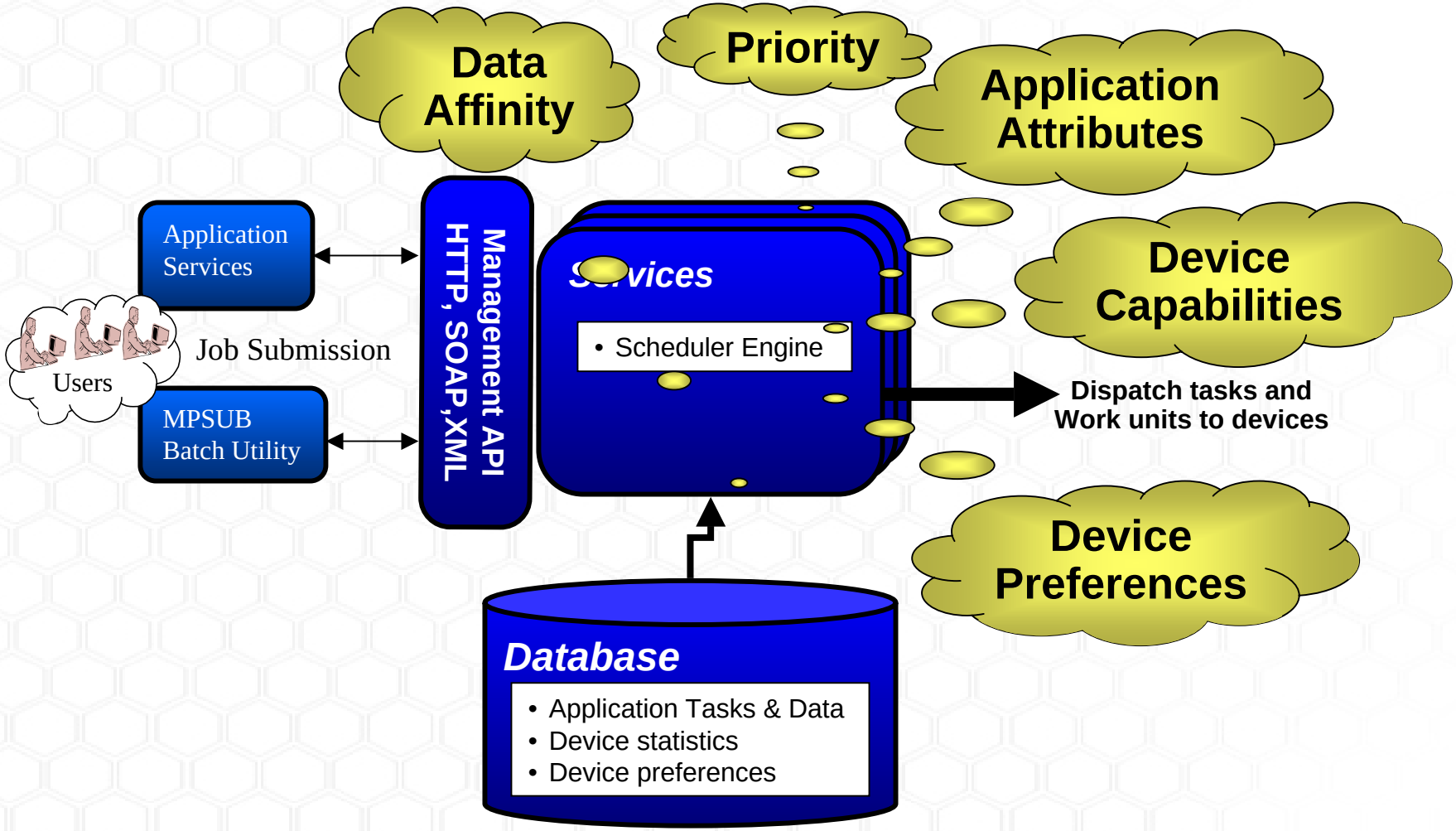
- **MetaProcessor Task API**
  - Optional source code modifications for task check-pointing and monitoring

- **MetaProcessor SDK**
  - Detailed documentation for all components
  - Reference implementation and value-added modules for rapid development

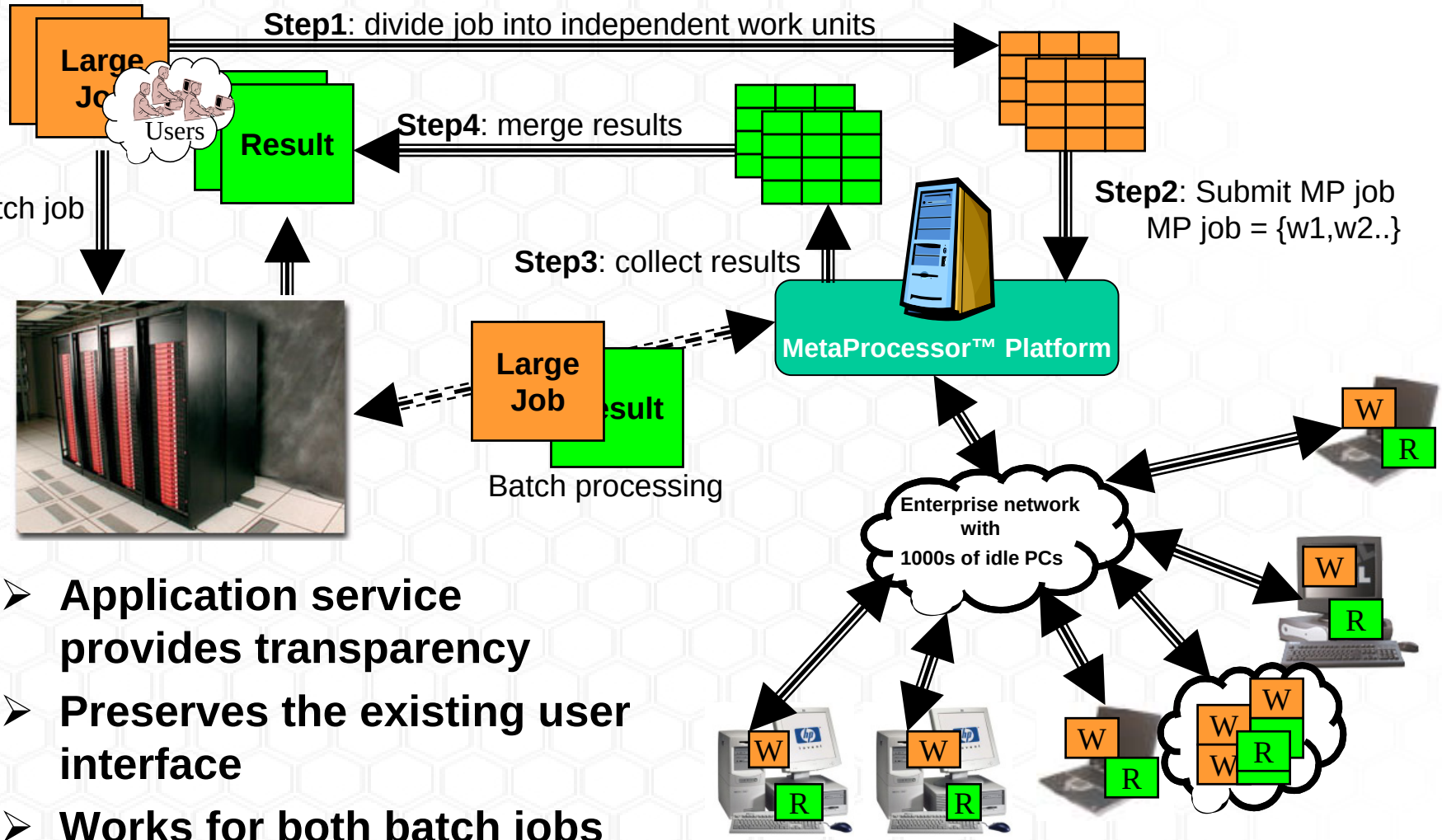
Data Management

- **Registration**
  - Data is registered as *work data* and *resident data* using MGSI
  - Data is uniquely named to avoid name collision
  - Data can be grouped as *work data set* and *resident data set*
- **Resident Data**
  - Data is always cached on devices to reduce network traffic
  - Data once registered may be shared by many jobs and users
  - Enables scheduling based on resident data affinity
- **Annotation**
  - Enables application services to store information such as, a data index

Workload and Resource Management



End-user Transparency



- **Application service provides transparency**
- **Preserves the existing user interface**
- **Works for both batch jobs and data parallel jobs**

Case Studies

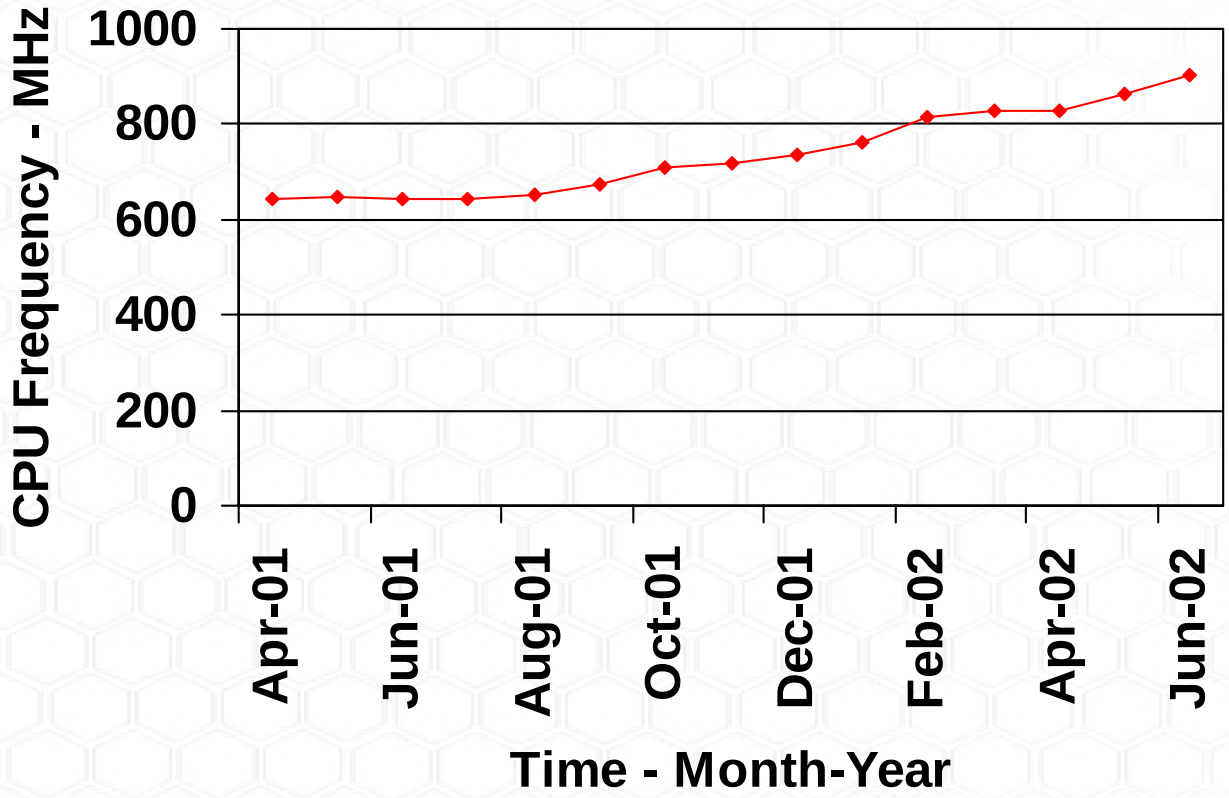
The Novartis Grid

- **Goal**
  - Accelerate lead identification and profiling 10X within one year
  - Find a way to accomplish the goal without raising costs 10X
- **Solution**
  - Adopt a PC-based Grid solution
- **Pilot Highlights**
  - Rapid deployment across 600+ Windows desktops
  - Exceeded expectations on all pilot criteria
  - Aggregated 3.18 years of processing time in first 7 days
  - Migrated proprietary application during the pilot
- **Results**
  - Currently running a 1000 device Grid
  - Response time on virtual screening reduced from days to hours
  - Network bandwidth usage increased by less than 2.5%
  - Investigating text mining application as well

Oxford University – A Global Grid

- **Goal**
  - Be able to screen a public database of small molecules against 12 protein targets identified in several cancers
- **Solution**
  - Adopt a Global Grid solution and use volunteered PCs
- **Project Highlights**
  - A total of over 1.6M PCs from around the world – 220 countries represented, including the Vatican
  - Unprecedented scope – project scope was increased 70-fold based on the available compute power
  - Total CPU time is over 126,000 years; 200+ years every day!
- **Results**
  - Other related projects completed: Anthrax screen in 24 days!
  - “Hits” will undergo the next stage of analysis
  - Largest, general-purpose, public compute Grid

The Global MetaProcessor Grid



Performance improved 50% in 14 months.

The asset appreciates over time!

Summary and the Future

- **The opportunity with enterprise grid computing (or P2P) is REAL**
- **There are some challenges but they can be overcome**
  - Applications and end-user transparency
  - Administration and manageability
- **The benefits are compelling and it is happening TODAY**
  - Novartis and Oxford University are example projects
- **The Future**
  - Every device is “plugged” into this type of grid
  - It becomes ubiquitous – you don’t have to think about it anymore
  - True peer-to-peer features enable less loosely-coupled applications