



Bioinformatic Pipelines a 10^g Perspective

Marcus Collins

Chief Database Technologist

Celera Genomics (Applera Corp)

4th Oracle Life Sciences User Group (OLSUG) Meeting
June 24th 2004



Introduction

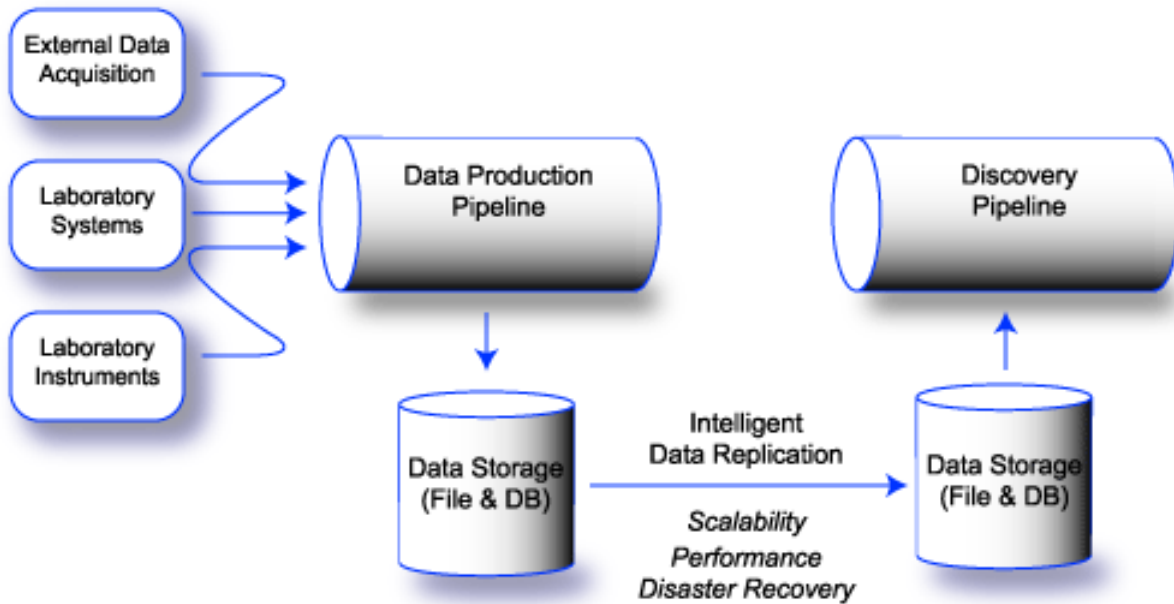
- Trends in Drug Discovery
- Bioinformatics Pipelines
- Issues
- Solutions (10g Perspective)
 - Hardware Clustering
 - Storage Area Networks
 - Administration and Monitoring
 - 10g New Features



Trends in Drug Discovery

- InSilico Drug Discovery
- Increased Automation
- Huge Increase in Data Volumes
- Human Genome Project (30,000 Genes)
- Protein Protein Interactions (500,000 Proteins)
- Personalized Medicine (Targeted Drug Discovery)
- Automation of Clinical Trials and Patient Records

Pipeline Architecture

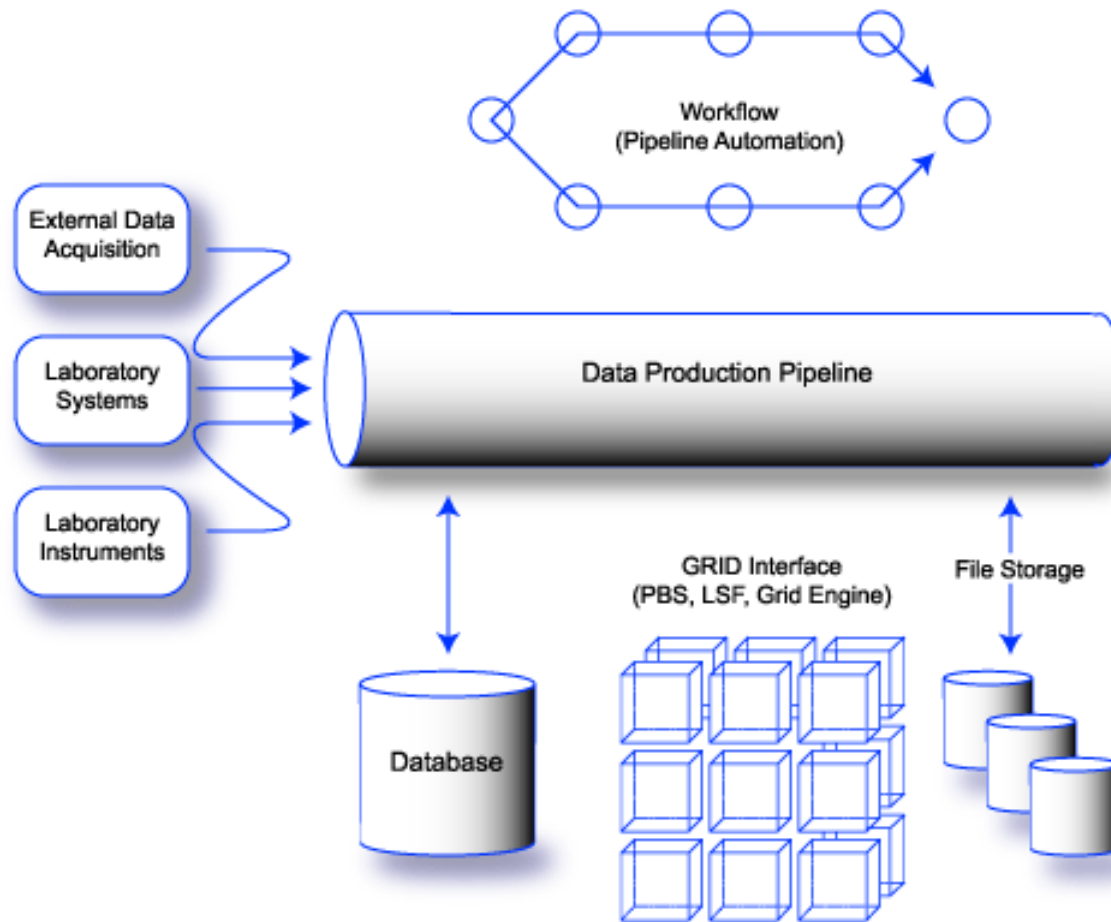




Key Features

- Simplified View (Coupling/Feedback Loop)
- Data Production
 - Runs at Instrument Speed
- Discovery
 - Runs at Human Speed

Data Production Pipeline

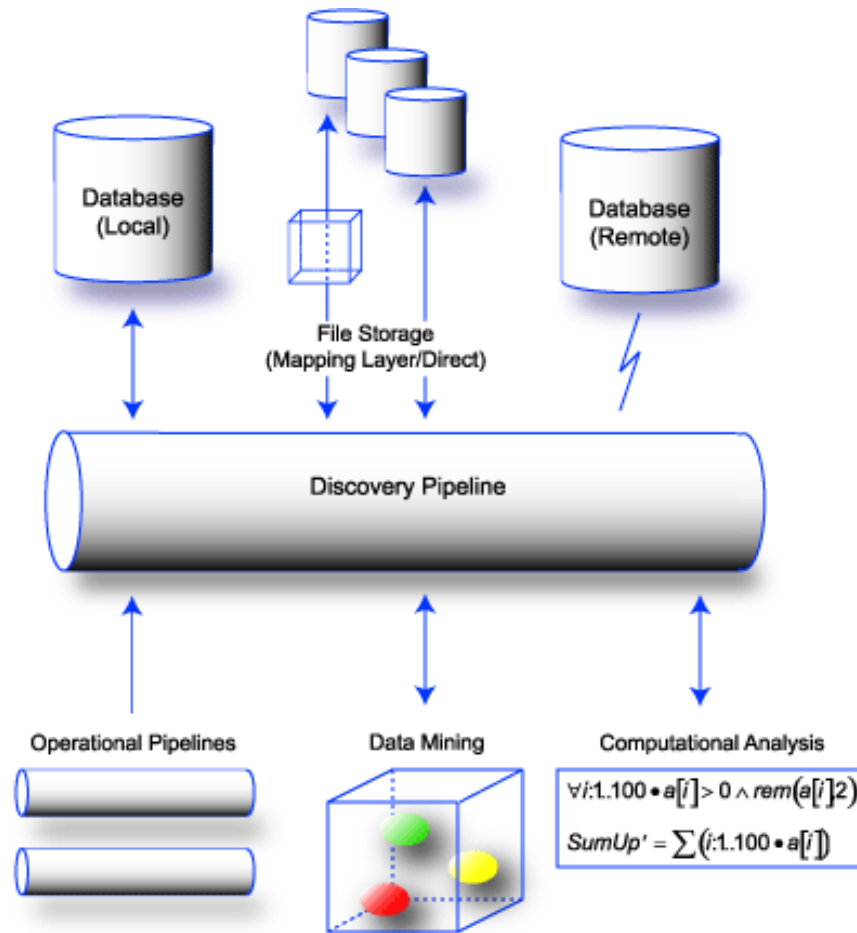




Key Features

- Database Design
 - OLTP & Data Warehouse (ETL & Summary)
- Instrument Integration
- Pipeline Control
 - Workflow and Process/Task Flow
- Compute Farm (Grid Engine) Integration
 - Tight Integration with Compute Engine
- Intelligent Data Replication
 - Automated Discovery

Discovery Pipeline





Key Features

- Heterogeneous Database Environment
 - View as Single Integrated Data Source
 - Searchable Flat Files (FASTA)
- Discovery Tools
 - Open Source Programming Tools
- Data Visualization
- Data Mining



Issues

- Reliability/Availability
 - Hardware and Storage
- Manageability
 - Multiple Database; Clustered Environment
- Data Movement
- Data Integration/Program Execution
- Backup Operations
- Compute Engine Integration



Oracle Database 10g Perspective

- Reliability/Availability
 - Linux Clusters
 - SAN/NAS
- Manageability
 - Oracle Common Management Infrastructure
- New Features/Enhancements
 - Data Pump
 - Data Integration/Program Execution
 - > Web Services/SOA
 - > XML DB
 - Recovery Manager (RMAN)
 - BLAST in the Database



Hardware Clustering

- Resource Failover
- Concurrent Access (Load Balancing)
- High Performance Compute Clusters
 - Massively Parallel Supercomputers/MPI
 - Used Only For Computational Purposes

Hardware Clustering

- Resource Failover
- Concurrent Access (Load Balancing)
- High Performance Compute Clusters
 - Massively Parallel Supercomputers/MPI
 - Used Only For Computational Purposes



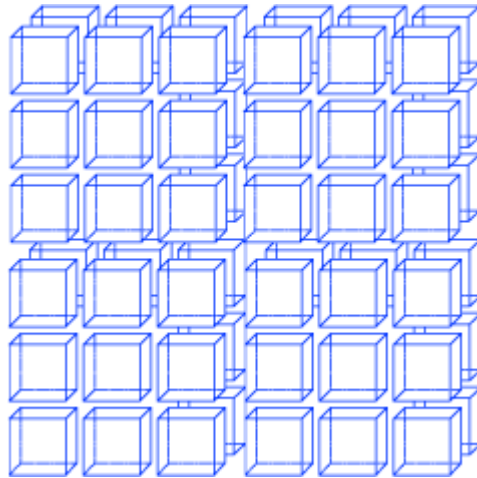
Hardware Clustering

- Resource Failover
- Concurrent Access (Load Balancing)
- High Performance Compute Clusters
 - Massively Parallel Supercomputers/MPI
 - Used Only For Computational Purposes



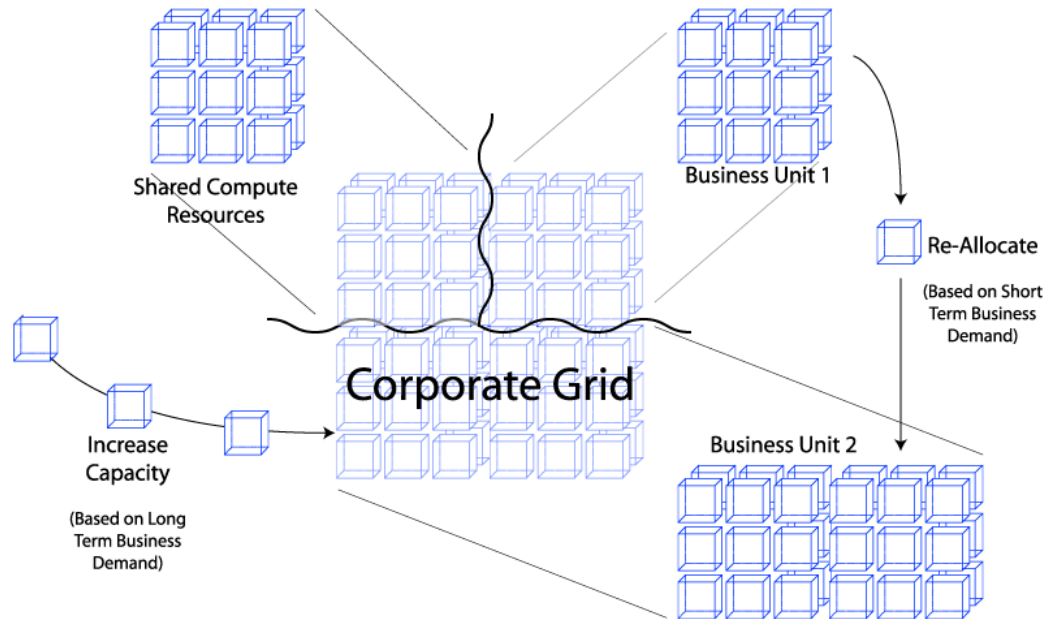
Hardware Clustering

- Resource Failover
- Concurrent Access (Load Balancing)
- High Performance Compute Clusters
 - Massively Parallel Supercomputers/MPI
 - Used Only For Computational Purposes



Hardware Clustering

- Convergence of Technologies
 - Grid Computing
 - Linux Clusters



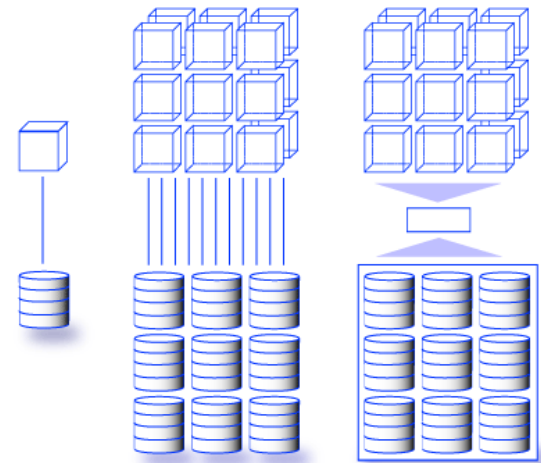


Hardware Clustering

- Benefits of Grid Computing
 - COTS
 - Scalability
 - Various High Availability Strategies
 - > Concurrent Access (RAC)
 - > Software Resource Failover
 - > Hardware Resource Failover (VMware VMotion)
- Current Issues
 - Cluster Ready Services (Based on Compaq TruCluster)
 - File System (Cluster File System)
 - > OCFS (Version 1; Version 2 Promised)
 - > ASM (Version 1; No Visibility at O/S)
 - > Raw Volumes (No Visibility at O/S)
 - > 3rd Party (PolyServe)
 - Cost (Full Loaded/Production Configuration)
 - Administration (Custom Configurations Increase Costs)
 - > True Cost Benefits of Grid Realized with Massive Standardization

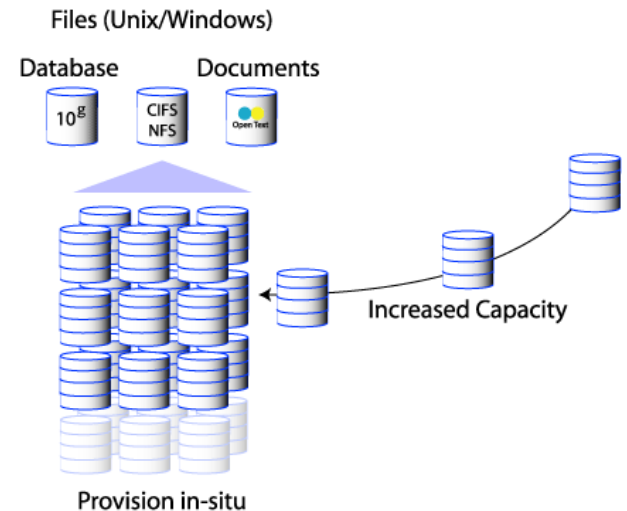
Storage Area Networks

- Large Data Volumes (Set to Increase)
- SAN Technology Maturing
- Consolidated Storage
 - Centralized Administration
 - Increased Performance
 - High Availability



Storage Area Networks

- Large Data Volumes (Set to Increase)
- SAN Technology Maturing
- Consolidated Storage
 - Centralized Administration
 - Increased Performance
 - High Availability
- Presentation On-Demand
 - SAN
 - NAS (Unix and Windows)
 - CAS (Information Lifecycle Management)





Storage Area Networks

- Implications for DBA's
 - No Easy Answer to “Where is my data?”
 - Storage Configuration
 - > Standard Allocation Units (e.g. 100Gb)
 - > Based on SLA (e.g. High, Low Performance)
- New Question
 - “How big a database can we build?”
 - Technology no Longer the Barrier
 - > Administration Tools
 - > DBA Skills



Administration

Evolving Role of the DBA

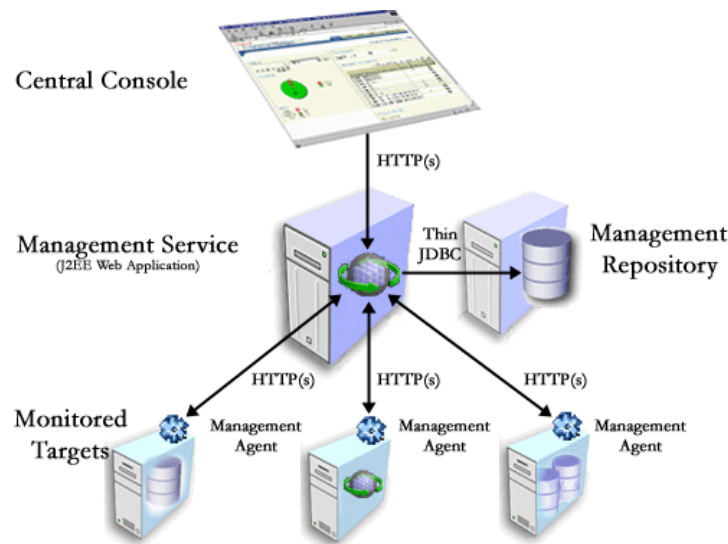
- Traditional Skills
 - Administration (Space, Backup, Data Loading etc.)
 - Performance Tuning
 - Schema Design

Oracle 10g Promises the “Self Managing Database”

- New Role (Emphasis on “Value Add”)
 - Business Knowledge
 - Information Lifecycle
 - > Information Flow Within the Organization
 - Database Design (Logical and Physical)
 - Infrastructure Design

Administration/Monitoring

- Enterprise Manager
 - Enhanced (DB, Hosts, iAS)
 - Concerned Over Reliability (9.2.0.3 – Stable)
- Automatic Workload Repository (AWR)
- Automatic Database Diagnostic Monitor (ADDM)





Administration/Monitoring

- Automatic Workload Repository (AWR)
 - Replacement for Statspack
 - Automatic Snapshot Collection
 - Repository for Advisors
- ADDM
 - Diagnostic Report After Each Snapshot
 - Performance Analysis Section of OEM
 - SQL Tuning



SQL Tuning

- Optimizer Requirements
 - Statistics/Statistics/Statistics
 - Less Inclined to “Flip” (Fail Gracefully)
 - More Accurate
 - Learns
- SQL Tuning Advisor
 - Automatic Statistics Collection
 - Optimizer Mode – “normal” or “tuning”
 - Create Tuning Task
 - > Recommendations (Additional Index)
 - > SQL Profile

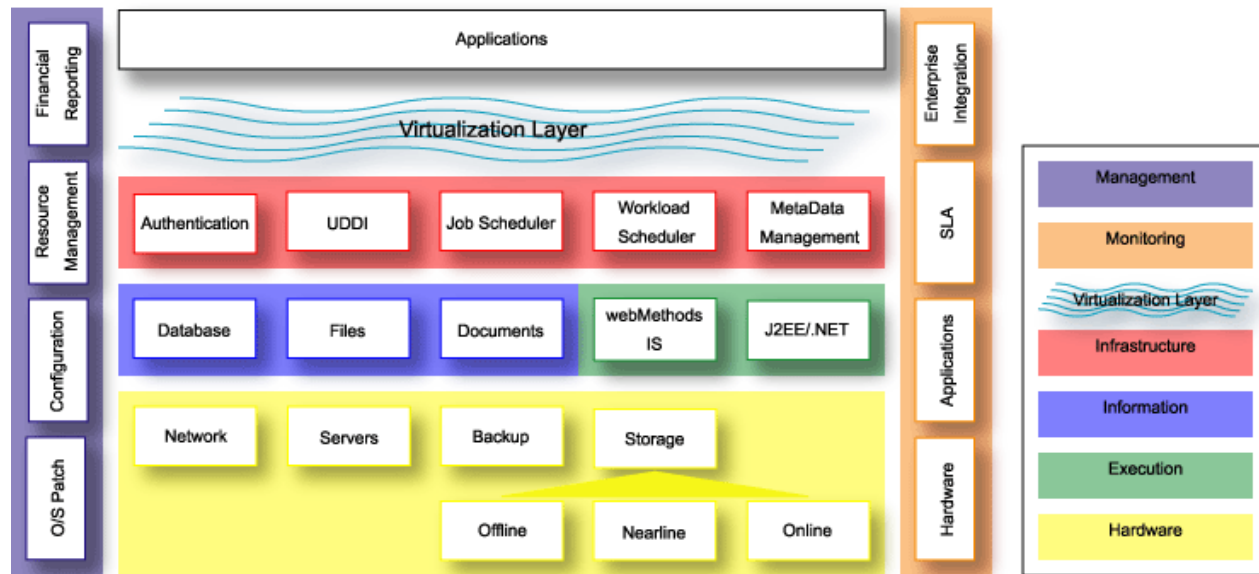


Data Pump

- expdp/impdp
- Two Modes
 - External Table (ET)
 - > RI Constraints/Global Indices
 - > Intra-Partition Parallelism
 - > PQ Slaves
 - Direct Path (DP)
 - > Faster
- Network Mode
 - Support for Named Pipes (Overlapping exp/imp)
- DDL Transforms
 - Remap Tablespace/Schema

Web Services/SOA

- Expose Database Through Web Services
 - PL/SQL (Web Services Call Out – UTL_DBWS)
 - PL/SQL or SQL (Expose via Web Services)





XML DB

- XML Integration Tool
- Issues with Schema Volatility
 - Persistent XML Store
 - Standards Based XML Only
- 10g XML Schema Evolution
 - Copy Documents to Temporary
 - Drop and Re-Register Schema
 - Transfer/XSLT Documents



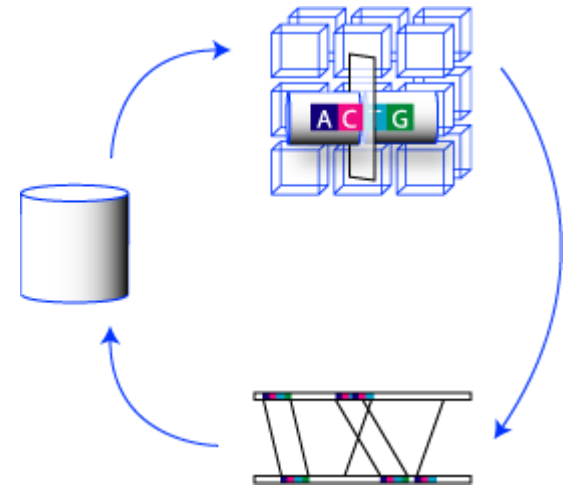
Recovery Manager

- Tool of Choice
- All Backups Online
 - No Requirement for Offline (EVER)
 - Online Supports Incremental Backups
- Issues Resolved in 10g
 - Layered Backup
 - > Incremental Performance
 - > Write Once/Read Many
 - > Block Change Tracking File
 - Performance
 - > Rate Limit (rate=nn)
 - > Backup Duration



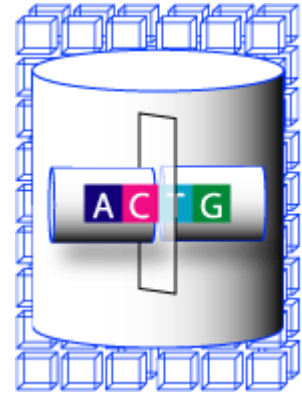
BLAST in the Database

- Current Mechanism
 - Extract/FormatDB
 - BLAST(n)
 - SQL*Load (Alignment/HSP)
- Benefits
 - Scalable (LSF/PBS)
 - Current Hardware Model
 - > Dedicated Database Servers
 - > Shared Compute Servers



BLAST in the Database

- Potential Mechanism
 - BLAST(n)
 - Load Alignment/HSP
- Issues
 - Performance
 - > Database vs. Compute Model
 - > Volume (10,000's)
 - Scalability
 - > Current Hardware Model
 - Dedicated Database Servers
 - > Future Hardware Model
 - Grid Model





Best of the Rest

- Regular Expressions
 - Can I have this in 9i?
- 10g Segment Management (ASSM)
 - Reduce Table Fragmentation
 - Instrument Black Box Database
- Transportable Tablespaces
 - Cross Platform Support
- Analytics



Summary

- Trends in Drug Discovery
- Bioinformatics Pipelines
- Issues
- Solutions (10g Perspective)
 - Hardware Clustering
 - Storage Area Networks
 - Administration and Monitoring
 - 10g New Features



Questions & Answers

marcus.collins@celera.com