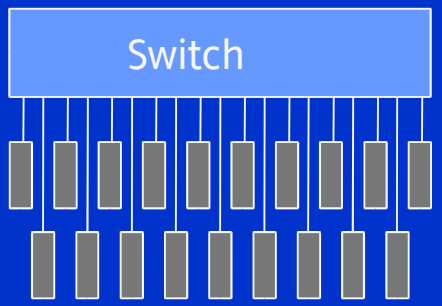
Aspects of a Processing Grid

Peter T. Kirstein P.Kirstein@cs.ucl.ac.uk
Søren-Aksel Sørensen S.Sorensen@cs.ucl.ac.uk
Stefano A. Street S.Street@cs.ucl.ac.uk
Sheng Jiang S.Jiang@cs.ucl.ac.uk
Department of Computer Science
University College London

Processing scenarios

- Batch processing
  - Large number of small (10's mins) jobs
  - Typical for Bioinformatics and HEP.
  - Resource hungry but not fussy.
  - Large data sets.
- Interactive experimentation
  - Single distributed job
  - User controlled (haptic steering)
  - Dynamic resource requirements
  - Performance sensitive (latency)
  - Visualization and steering

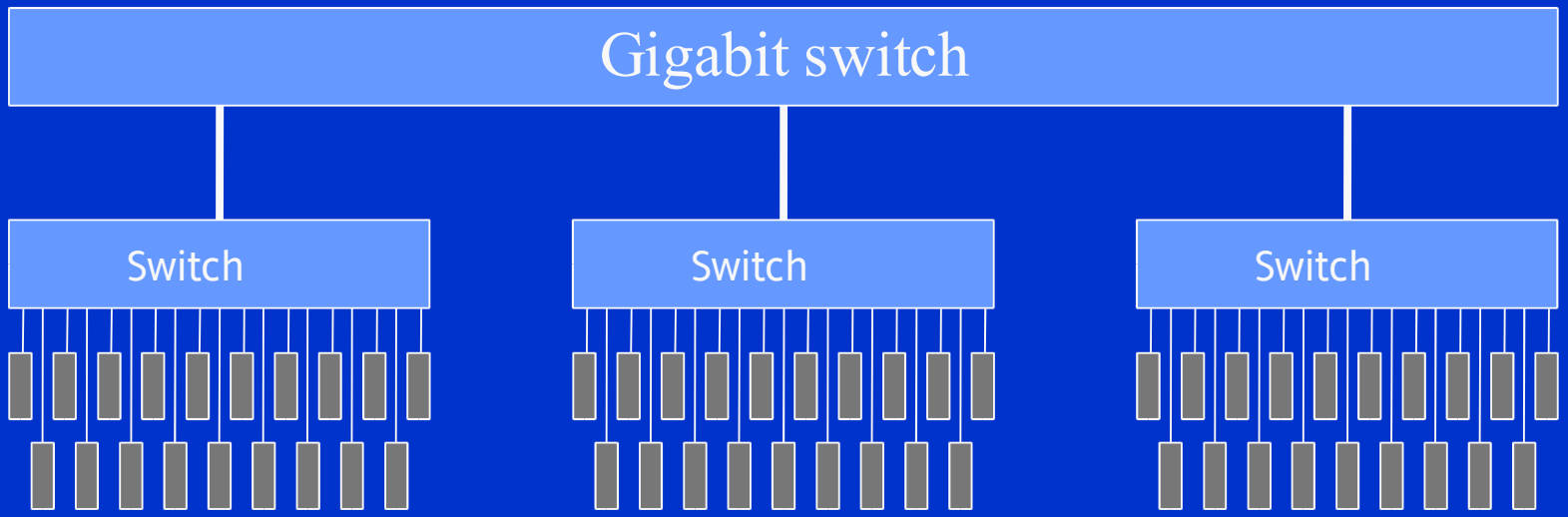
Cluster computing



- Advantages
  - Cost effective solution to resource shortage
  - Simple security (isolated environment)
  - Easy management.
- Problems
  - Switching cost (Ethernet)
  - Task mixing (delay vs. throughput)
  - Resource management (time sharing, performance protection)

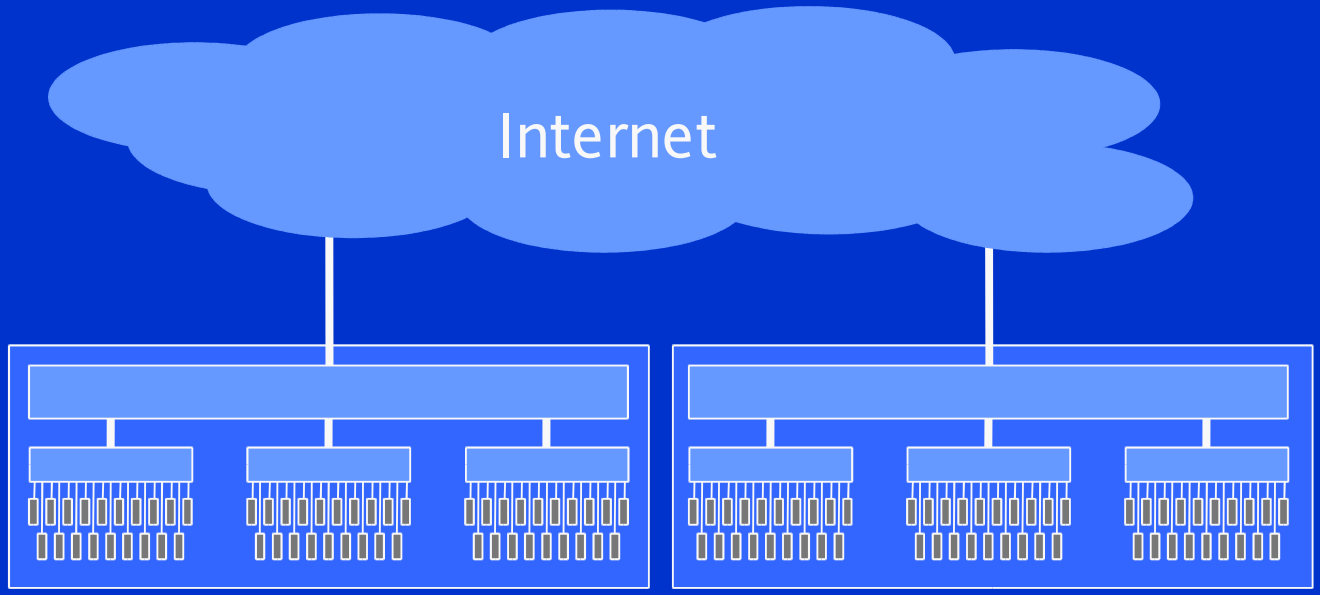
Cluster of clusters

- Large systems can be created from a hierarchy of sub-clusters.
- The UCL cluster uses this architecture

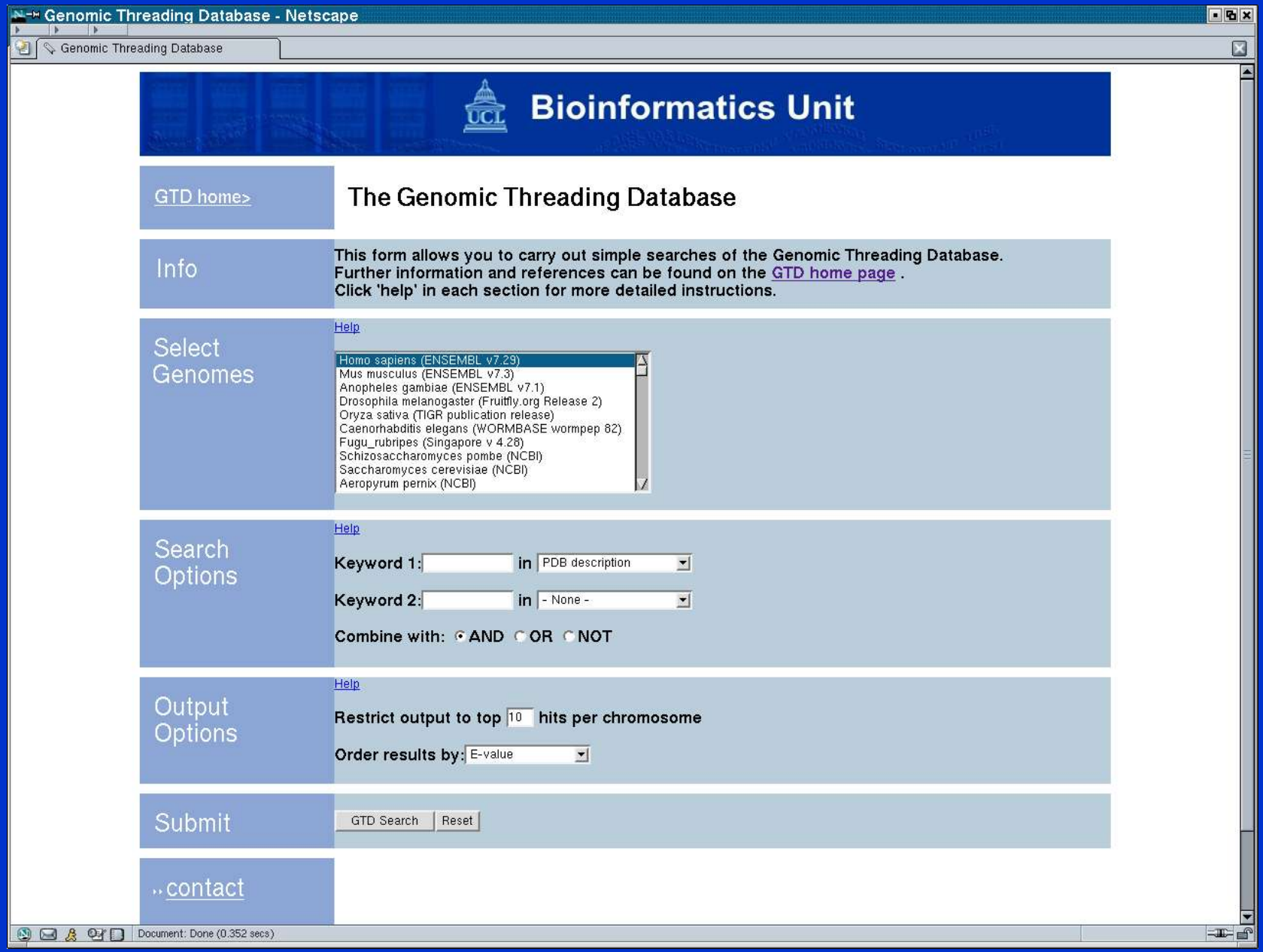


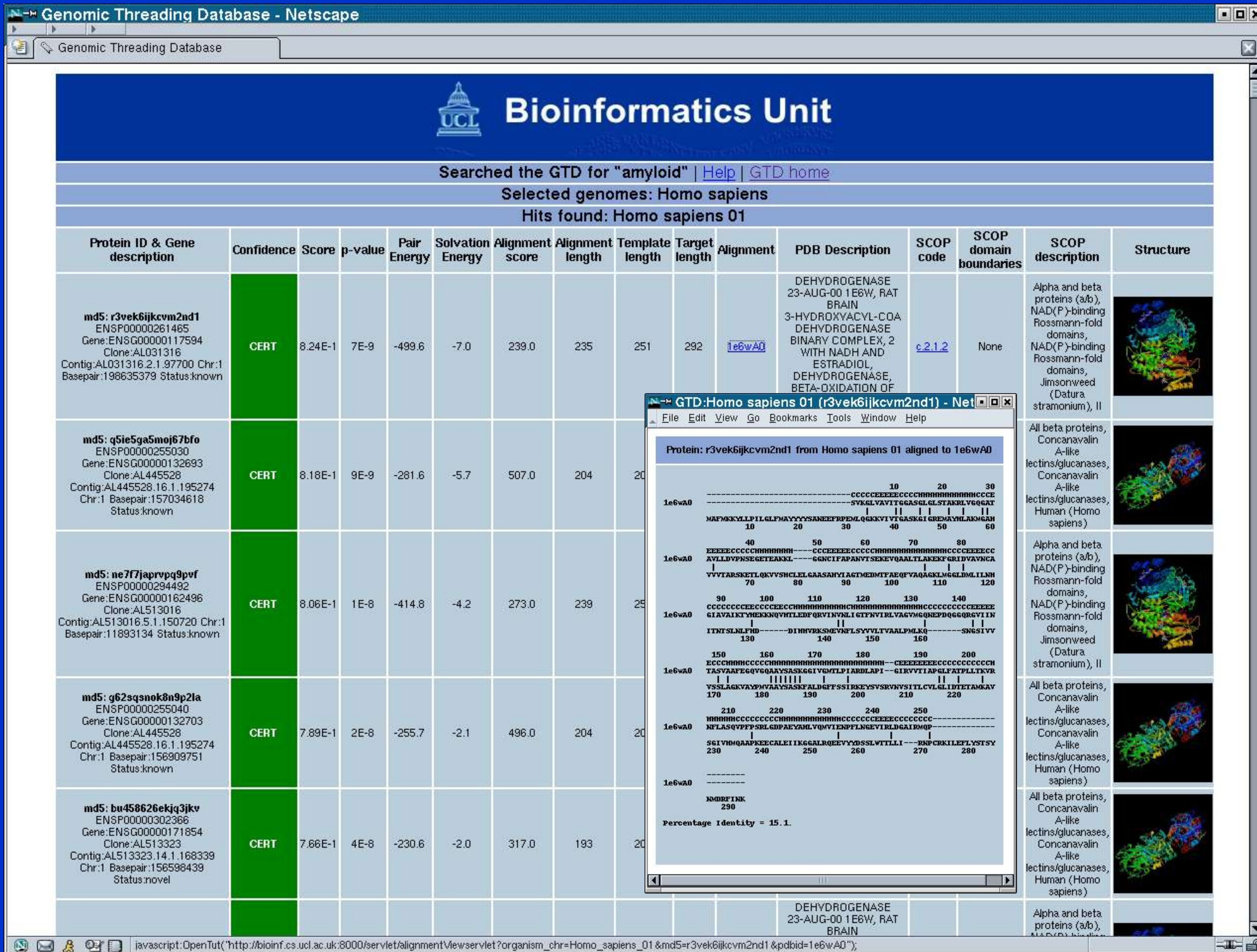
- Still growth problems (eg. power and cooling)
- Utilisation problems

Multi-domain clusters



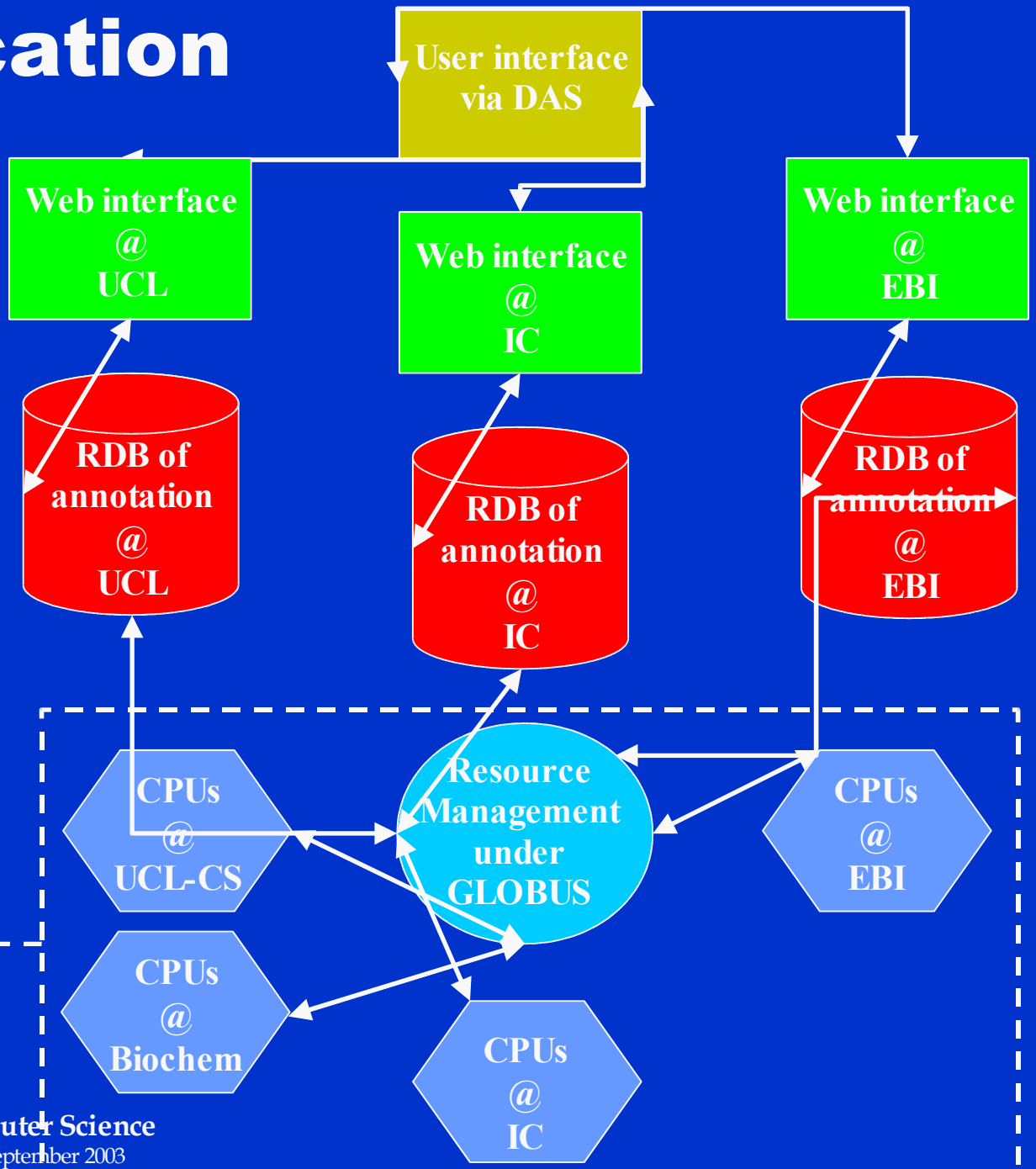
- Problems
  - Security
  - Load balancing
  - Resource discovery


Genomic Threading Database - Netscape
Genomic Threading Database
Bioinformatics Unit
GTD home>
The Genomic Threading Database
Info
This form allows you to carry out simple searches of the Genomic Threading Database.
Further information and references can be found on the GTD home page .
Click 'help' in each section for more detailed instructions.
Select Genomes
Help
Homo sapiens (ENSEMBL v7.29)
Mus musculus (ENSEMBL v7.3)
Anopheles gambiae (ENSEMBL v7.1)
Drosophila melanogaster (Fruitfly.org Release 2)
Oryza sativa (TIGR publication release)
Caenorhabditis elegans (WORMBASE wormpep 82)
Fugu_rubripes (Singapore v 4.28)
Schizosaccharomyces pombe (NCBI)
Saccharomyces cerevisiae (NCBI)
Aeropyrum pernix (NCBI)
Search Options
Help
Keyword 1: in PDB description
Keyword 2: in - None -
Combine with: AND OR NOT
Output Options
Help
Restrict output to top 10 hits per chromosome
Order results by: E-value
Submit
GTD Search
Reset
..contact
Document: Done (0.352 secs)



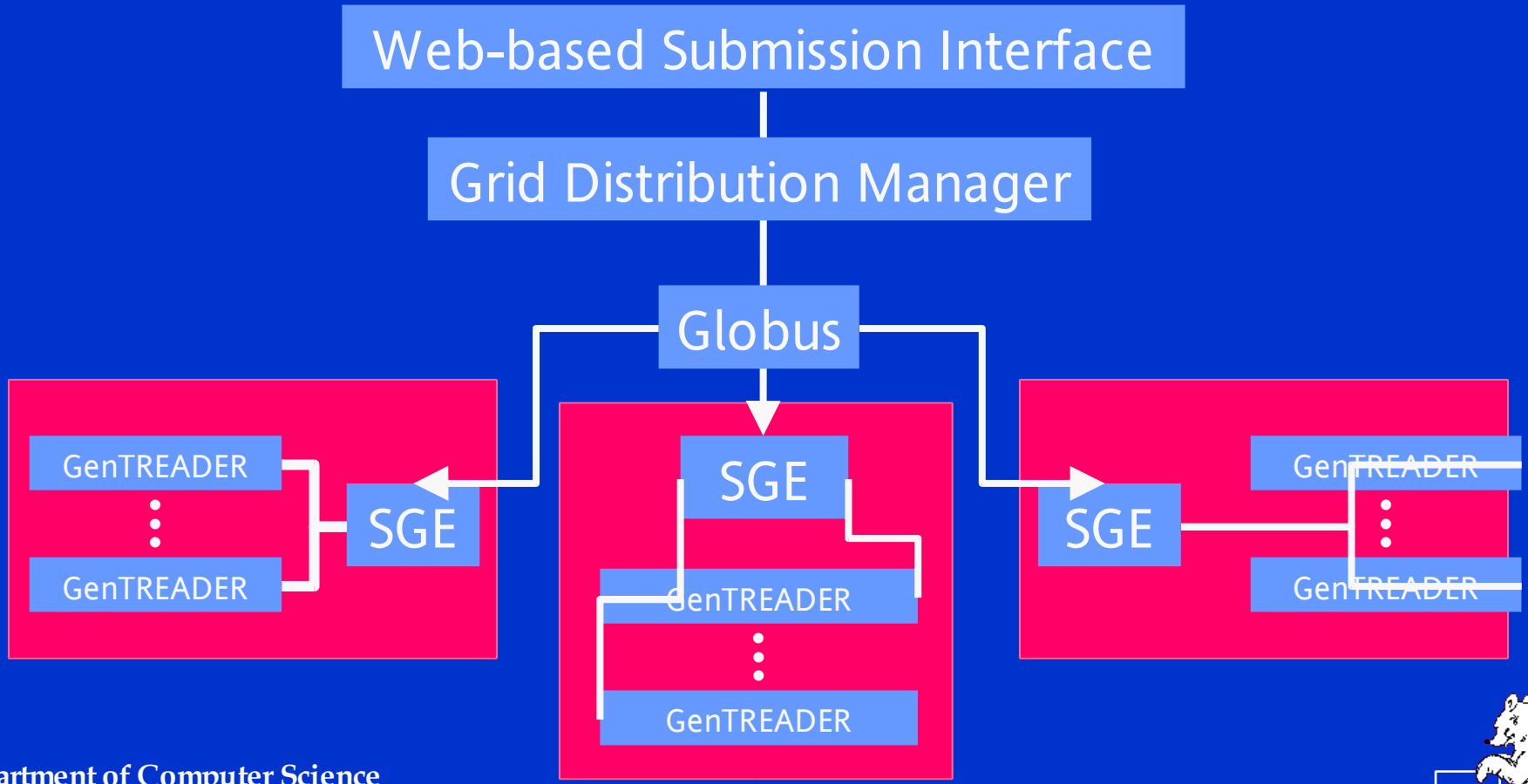
Protein ID & Gene description	Confidence	Score	p-value	Pair Energy	Solvation Energy	Alignment score	Alignment length	Template length	Target length	Alignment	PDB Description	SCOP code	SCOP domain boundaries	SCOP description	Structure
md5: r3vek6ijkcvm2nd1 ENSP00000261465 Gene:ENSG00000117594 Clone:AL031316 Contig:AL031316.2.1.97700 Chr:1 Basepair:198635379 Status:known	CERT	8.24E-1	7E-9	-499.6	-7.0	239.0	235	251	292	1e6wA0	DEHYDROGENASE 23-AUG-00 1E6W, RAT BRAIN 3-HYDROXYACYL-COA DEHYDROGENASE BINARY COMPLEX, 2 WITH NADH AND ESTRADIOL, DEHYDROGENASE, BETA-OXIDATION OF	c.2.1.2	None	Alpha and beta proteins (a/b), NAD(P)-binding Rossmann-fold domains, NAD(P)-binding Rossmann-fold domains, Jimsonweed (Datura stramonium), II	
md5: q5ie5ga5moj67bfo ENSP00000255030 Gene:ENSG00000132693 Clone:AL445528 Contig:AL445528.16.1.195274 Chr:1 Basepair:157034618 Status:known	CERT	8.18E-1	9E-9	-281.6	-5.7	507.0	204	20						All beta proteins, Concanavalin A-like lectins/glucanases, Concanavalin A-like lectins/glucanases, Human (Homo sapiens)	
md5: ne7f7japrvpq9pvf ENSP00000294492 Gene:ENSG00000162496 Clone:AL513016 Contig:AL513016.5.1.150720 Chr:1 Basepair:11893134 Status:known	CERT	8.06E-1	1E-8	-414.8	-4.2	273.0	239	25						Alpha and beta proteins (a/b), NAD(P)-binding Rossmann-fold domains, NAD(P)-binding Rossmann-fold domains, Jimsonweed (Datura stramonium), II	
md5: g62sqsnok8n9p2la ENSP00000255040 Gene:ENSG00000132703 Clone:AL445528 Contig:AL445528.16.1.195274 Chr:1 Basepair:156909751 Status:known	CERT	7.89E-1	2E-8	-255.7	-2.1	496.0	204	20						All beta proteins, Concanavalin A-like lectins/glucanases, Concanavalin A-like lectins/glucanases, Human (Homo sapiens)	
md5: bu458626ekjq3jkv ENSP00000302366 Gene:ENSG00000171854 Clone:AL513323 Contig:AL513323.14.1.168339 Chr:1 Basepair:156598439 Status:novel	CERT	7.66E-1	4E-8	-230.6	-2.0	317.0	193	20						All beta proteins, Concanavalin A-like lectins/glucanases, Concanavalin A-like lectins/glucanases, Human (Homo sapiens)	
											DEHYDROGENASE 23-AUG-00 1E6W, RAT BRAIN			Alpha and beta proteins (a/b),	

Application

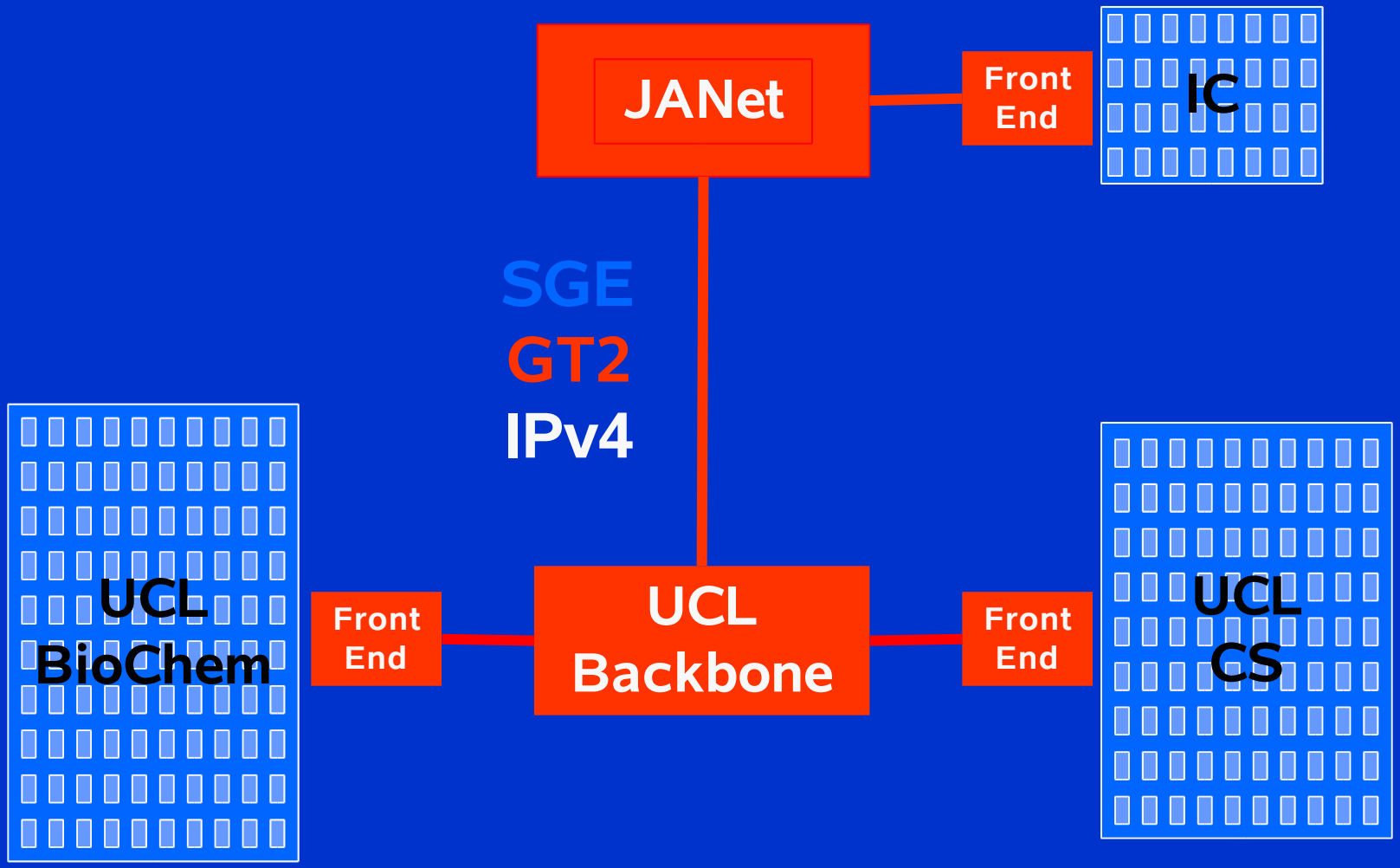


Distributed GenTHREADER Configuration

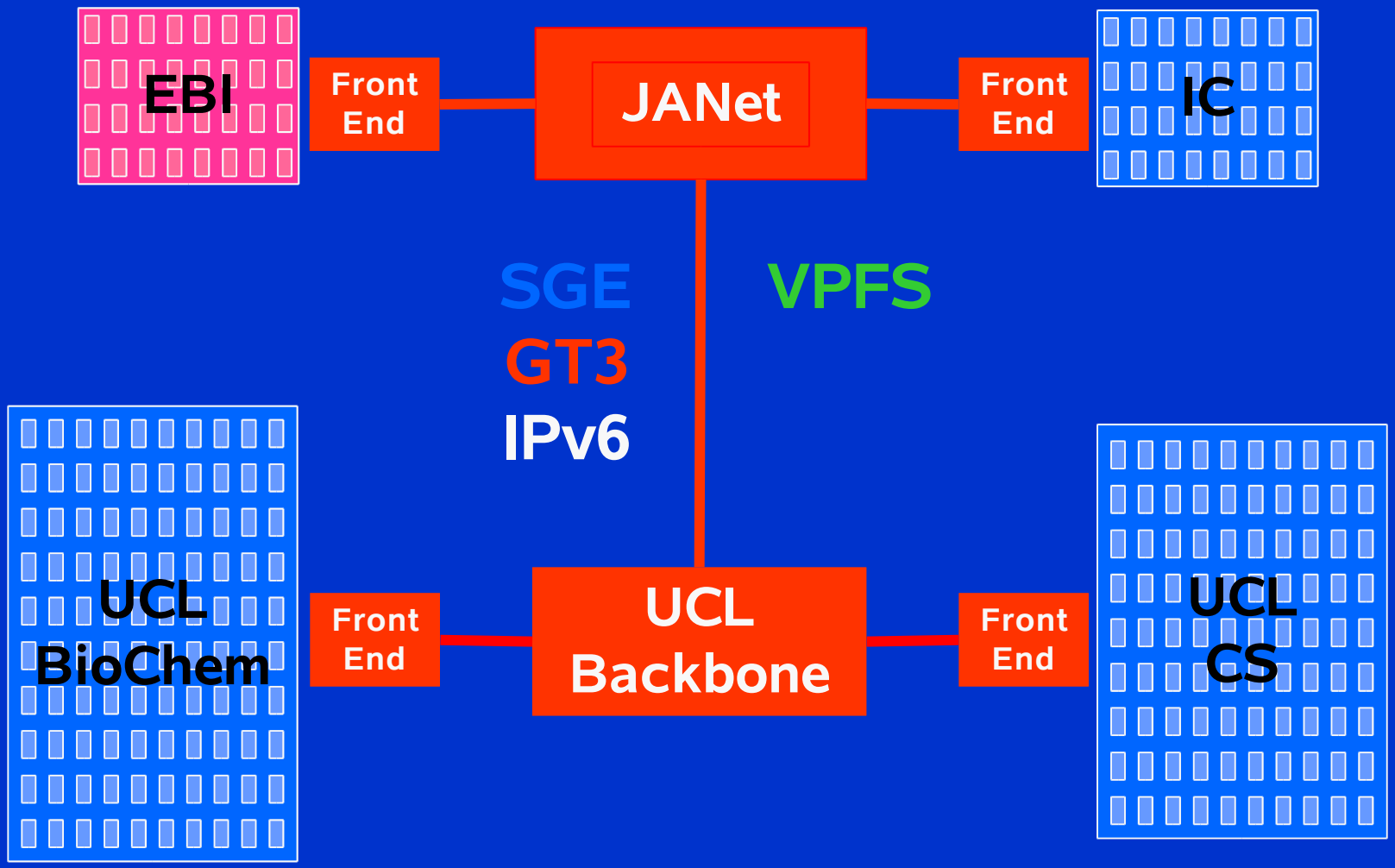
- User ‘annotation’ submission committed via Web Form
- Web service filters & divides sequence file into sub jobs



eProtein current production setup



eProtein future setup



IPv6 Advantages

- Larger Address Space
- Mobility Support
- Built-in Security
- Hooks for QoS
- Better Auto-configuration
- End-to-end Addressing
- Deploys New Technologies

Grid over IPv6

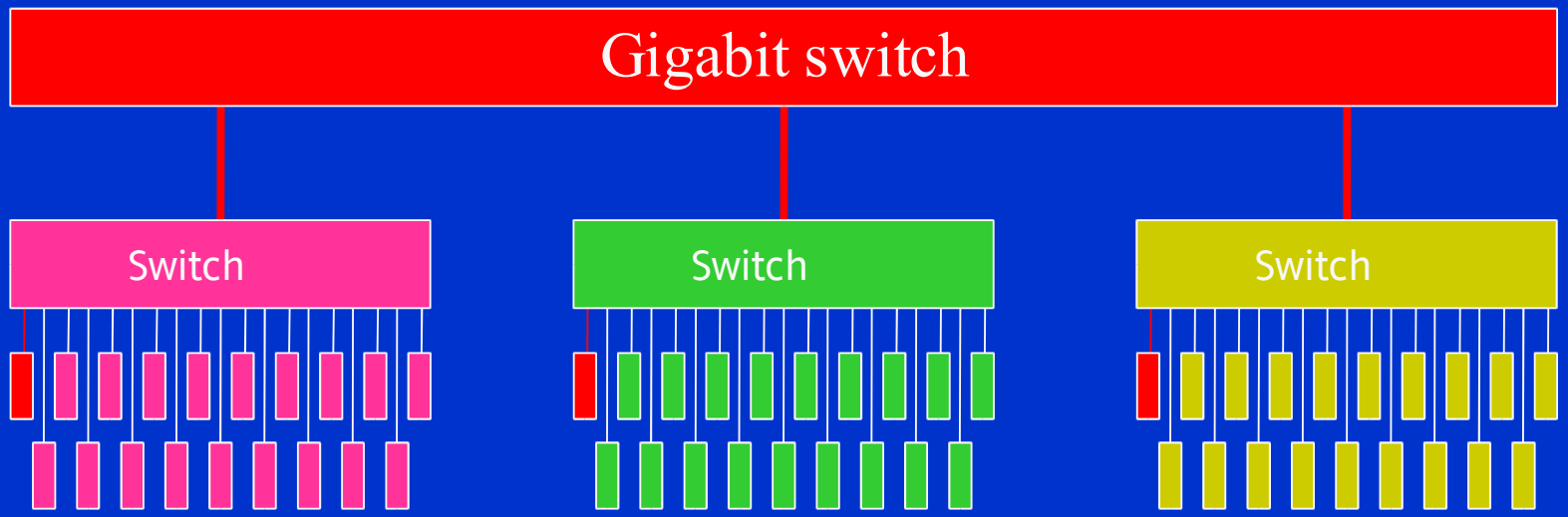
- Full availability of IPv6 will allow provision of better Grid services
  - Mobility support in distributed networks or global-scale networks
  - Performance potentially much better
  - Auto-configuration for resource discovery
  - Allow separation of addressing and routing
  - Better group communication
  - Better security

Demonstrator Setup

- Intel
- Linux (Red Hat 8 – 2.4.18)
- GT3 (alpha)
- Java (Sun SDK 1.4.1)
  - IP-independent - Class InetAddress used everywhere
    - Inet4Address
    - Inet6Address
  - Tomcat “lightweight edition”

Testing

- GenTREADER setup
- GT3 GRAM services
- Local subnets



Interactive use

- Distributed systems use PVM/MPI2
- Dynamic scheduling through dummies
  - Dummy process scheduled by SGE.
  - Extra resource communicated to application.
  - Dummy replaced by application process.
- Priorities used to favour interactive processes
- Cluster wide process control.