
KISTI's Activities on the NA4 Biomed Cluster

**Soonwook Hwang, Sunil Ahn,
Jincheol Kim, Namgyu Kim
and Sehoon Lee**

KISTI e-Science Division

Outline

- **DrugScreeener-G:**
 - Toward Integrated Environment for Grid-enabled Large-scale Virtual Screening
- **Data challenge on Diabetes using the newly developed performance-enhanced WISDOM Production environment**
- **New Developments in AMGA**

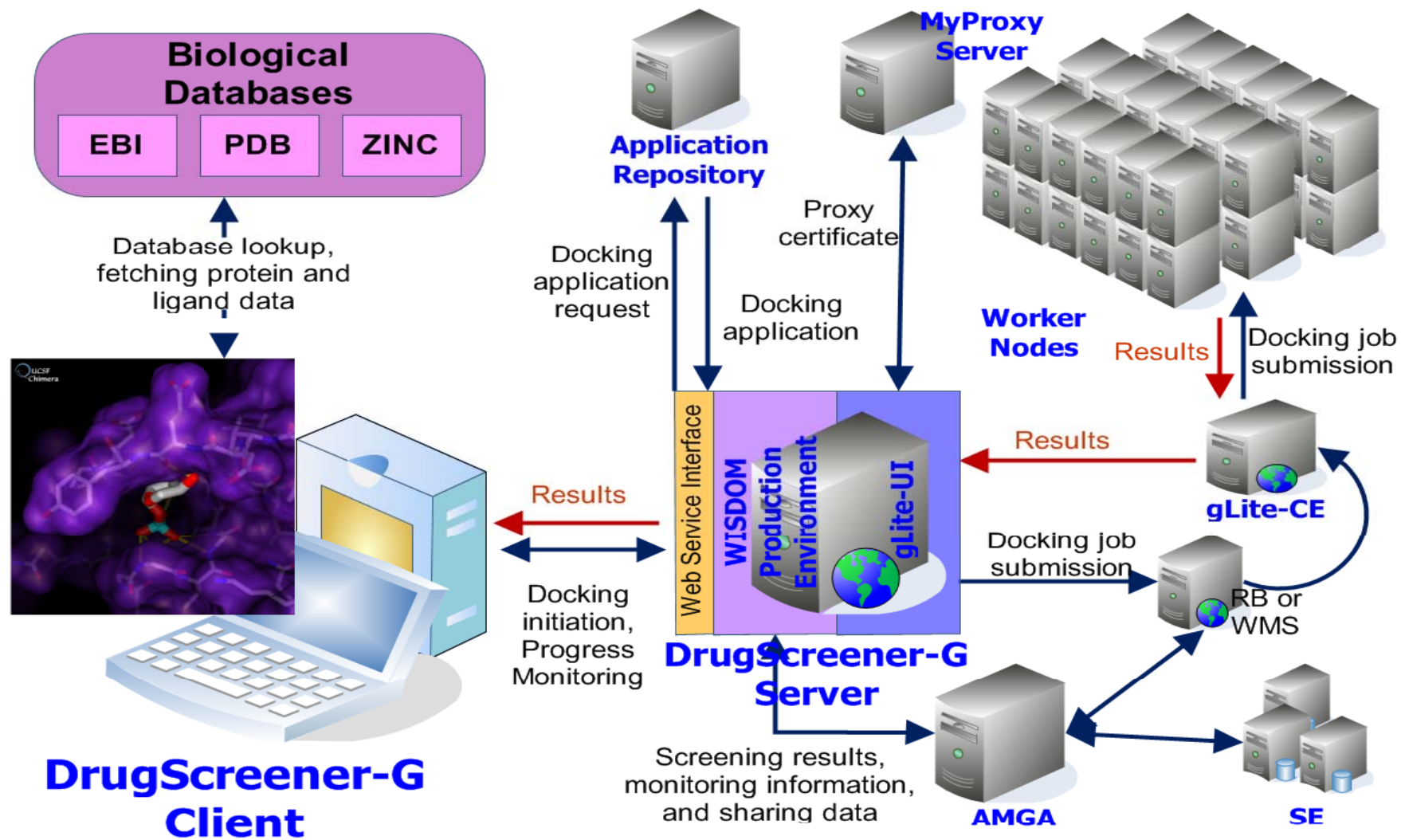
DrugScreener-G: Toward Integrated environment for Grid-enabled large-scale Virtual Screening

DrugScreener-G

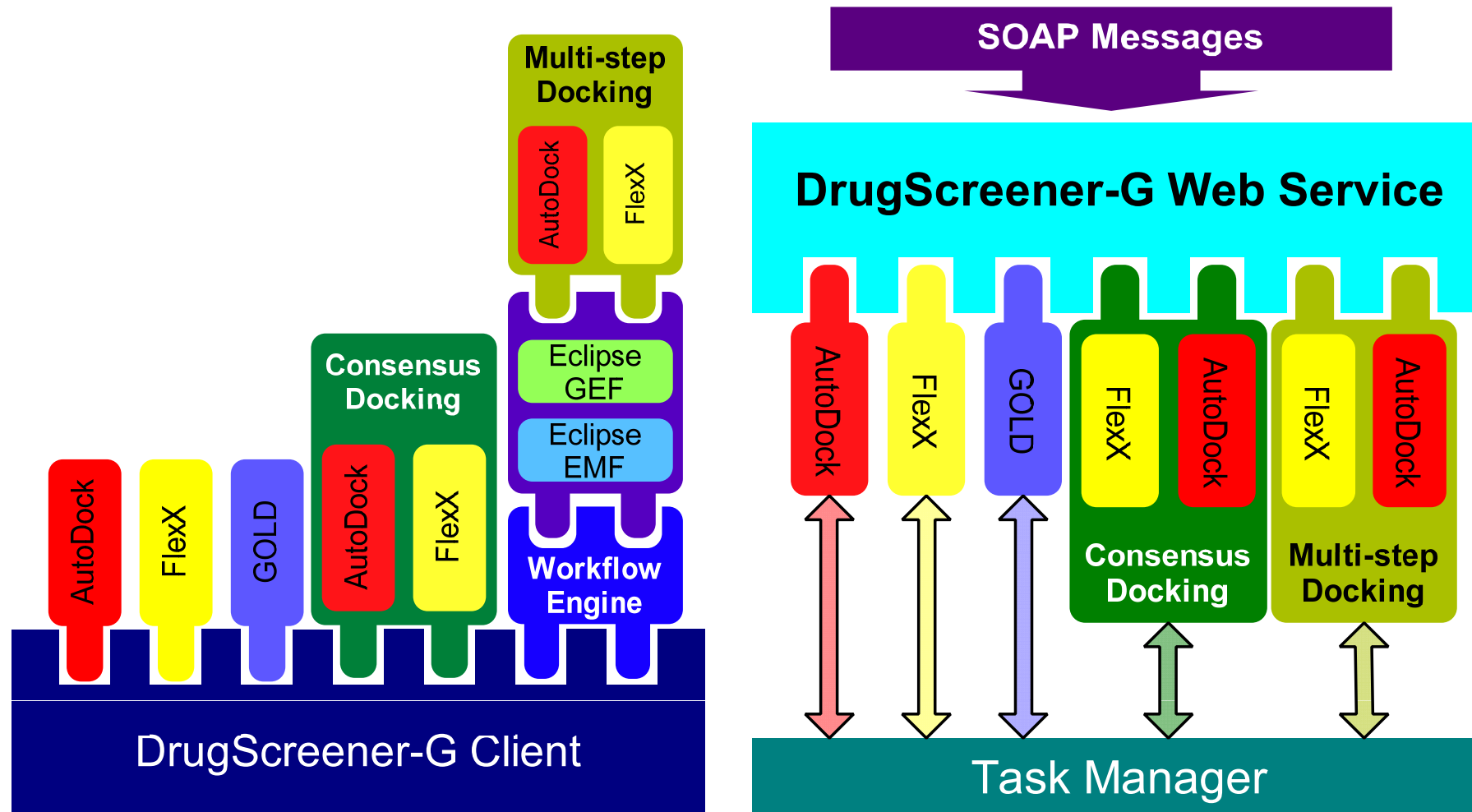
- **Objective**

- Providing a user-friendly integrated environment for Grid-based large-scale virtual screening for users without much knowledge of Grid computing to exploit full power of Grid computing infrastructure for drug discovery
- Target users: Bioinformaticians, Biologists, Drug Chemists

Virtual Screening in DrugScreeener-G



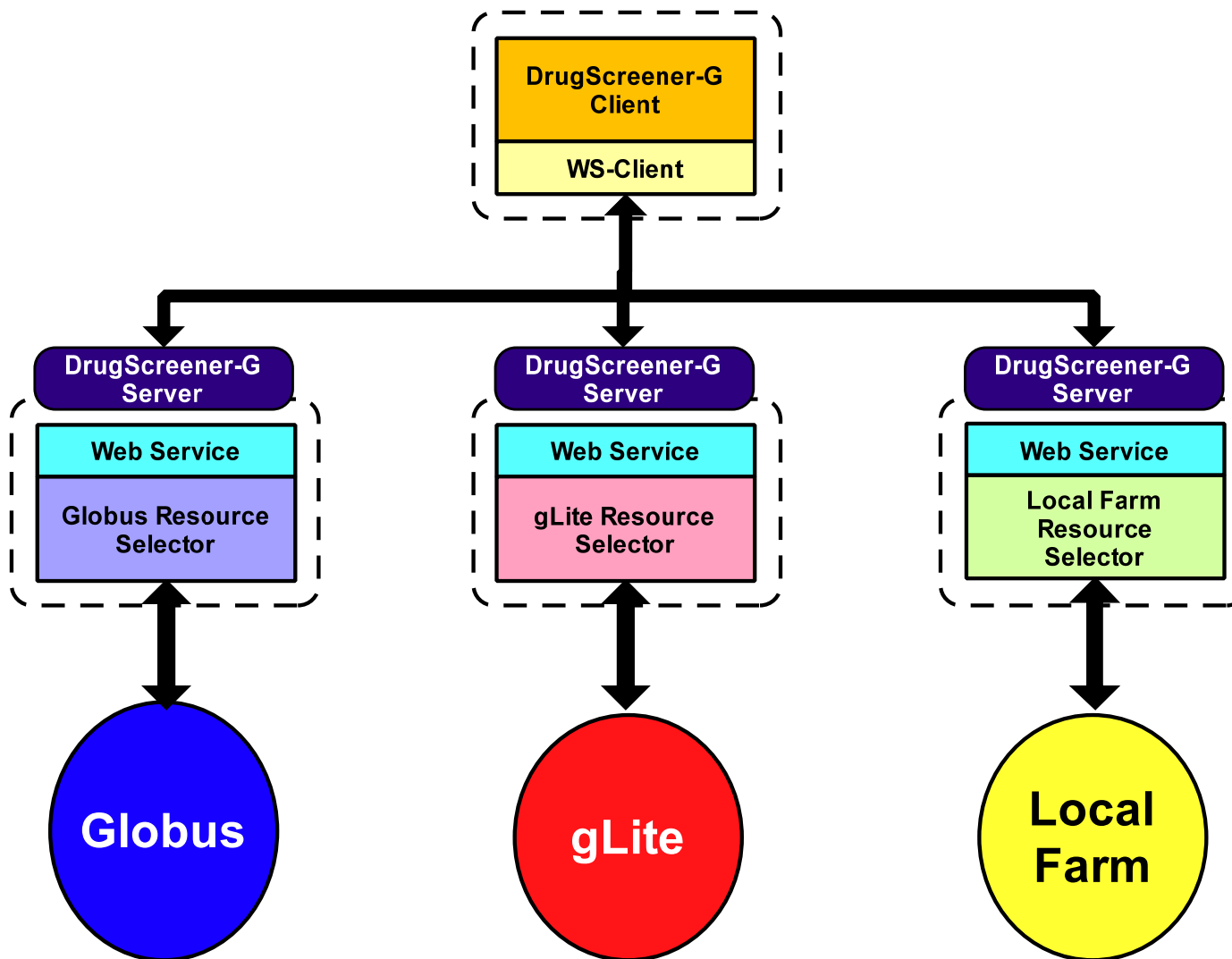
Plug-in Architecture in DrugScreener-G



■ Plug-ins on client's side

■ Plug-ins on server's side

Support for multiple platforms



Data Challenge on Diabetes

Data Challenge against diabetes type 2

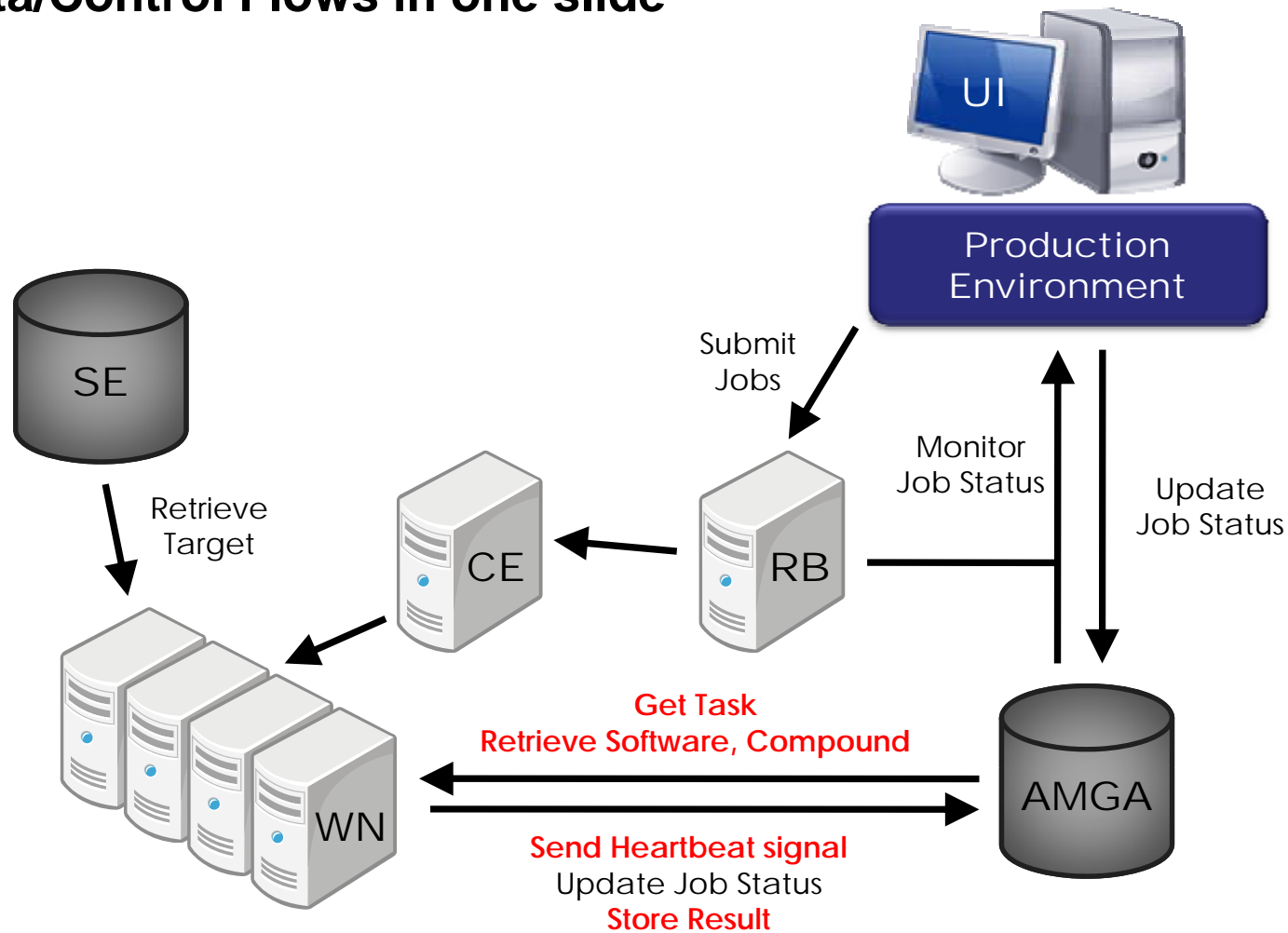
- Data challenge to Human pancreatic amylase inhibitor (1u2y), a target protein of diabetes type 2, with 308310 chemical compounds
- Achieved significant improvement in efficiency and throughput

	WISDOM-II	DIANE	KISTI
Total number of dockings	156,407,400	308,585	308,310
Estimated duration on 1 CPU	413 years	16.7 years	39.0 years
Duration of the experiment	76 days	30 days	2.4 days
Cumulative number of Grid jobs	77,504	2,580	103,583
Maximum number of concurrent CPUs	5000	240	7,370
Number of used Computing Elements	98	36	127
Crunching Factor	1983	203	5937
Distribution Efficiency	39%	84%	81%

Performance Enhancement in the WISDOM Production Environment

Evolution of WISDOM Production Environment

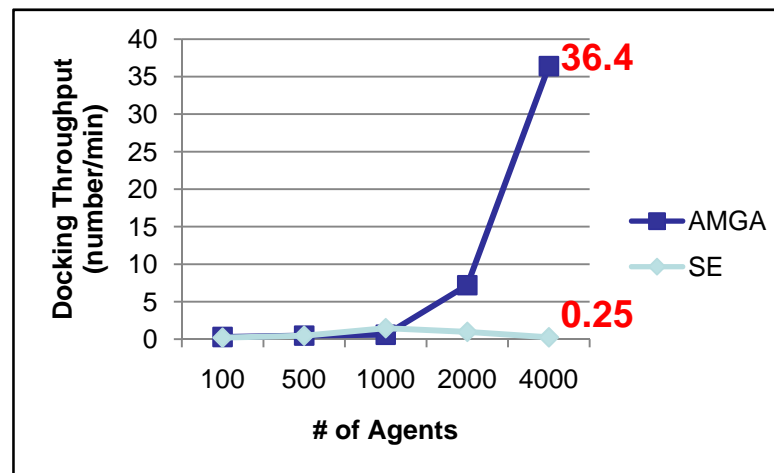
- Data/Control Flows in one slide



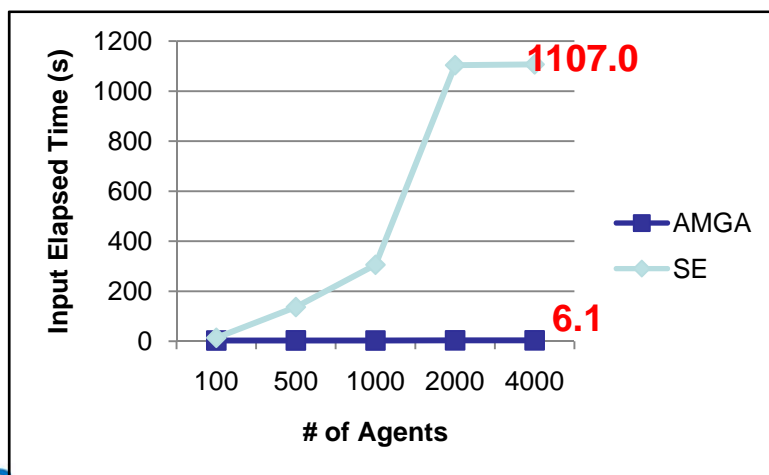
The exploitation of AMGA for storing ligands and docking results

- Why is AMGA used for the input/output data storing purpose?
 - AMGA performs way much better than SE in terms of throughput and latency

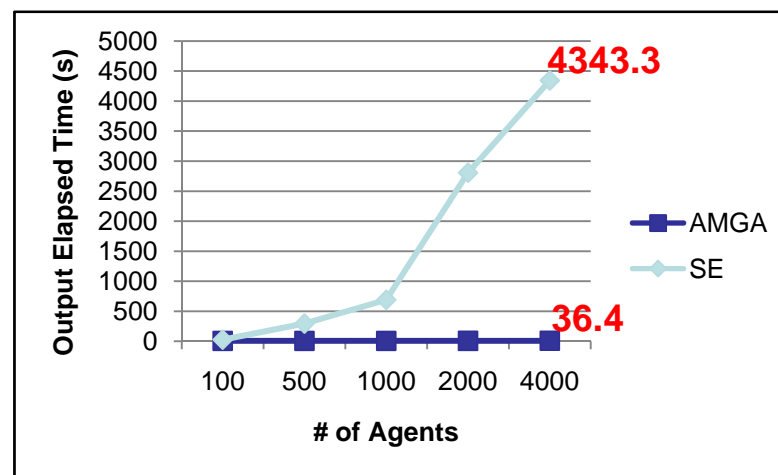
Docking Throughput



Time taken to retrieve a ligand



Time taken to store a docking result



The use of AMGA for task distribution purpose in the pull model of WISDOM PE

- A new AMGA operation is designed to allow jobs to fetch a task from the AMGA server in a single operation, rather than calling
- the AMGA source code modified and a new operation “updateAttr_single” added
- The throughput of task retrieval using the new operation is 70 times higher than that of using the existing AMGA operations

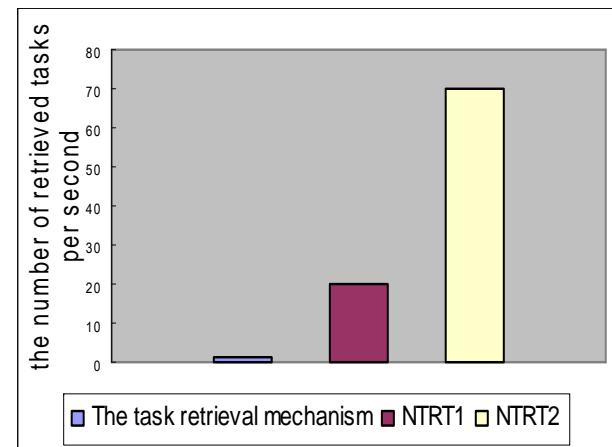
Job (Worker Node)

```
updateAttr_single task status "running" 'task:status = "waiting"'  
getAttr (entryName) task_id simulation_id
```

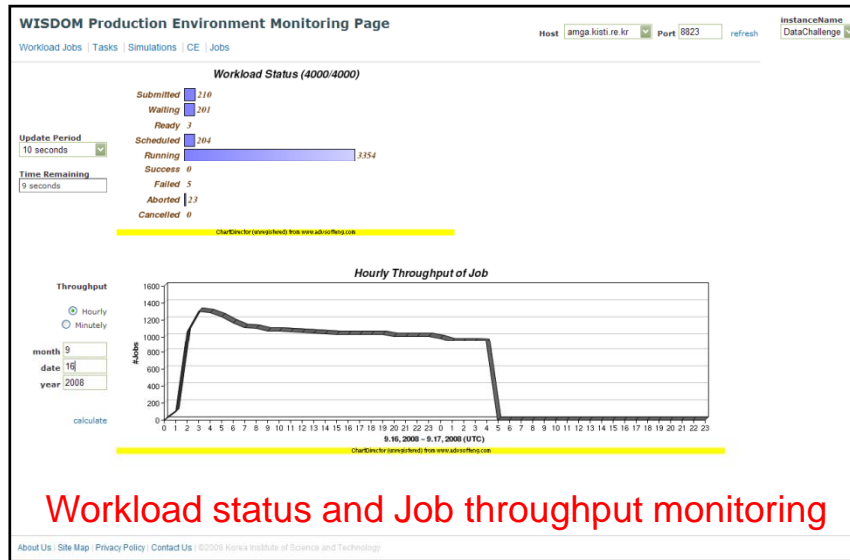
operation call

The “updateAttr_single” Operation (AMGA Server)

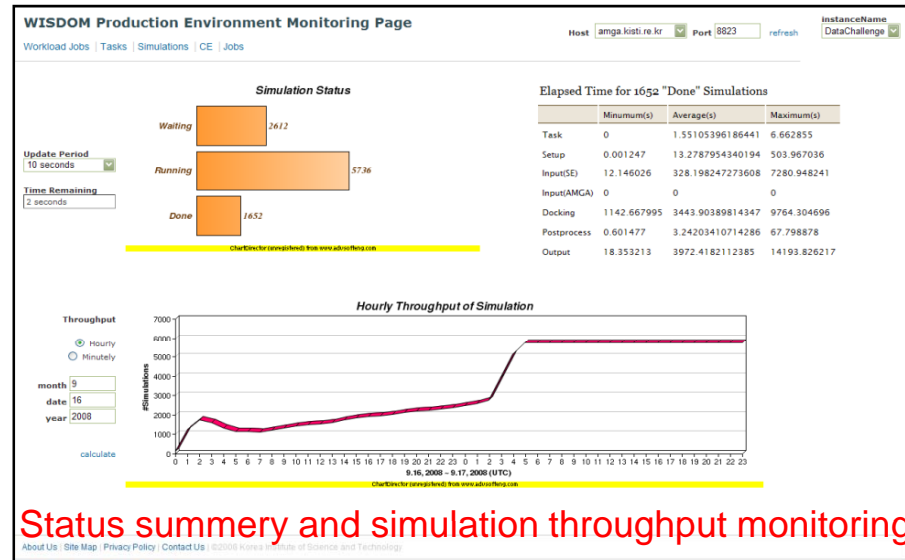
1. sem_wait(semaphore)
2. transaction
3. selectAttr min(task.task_id) 'task:status="waiting"'
4. updateAttr task status "running" 'task:task_id = (task_id)'
5. commit
6. sem_post(semaphore)
7. returns entry name



Enhanced Monitoring & Bookkeeping



Workload status and Job throughput monitoring



Status summary and simulation throughput monitoring

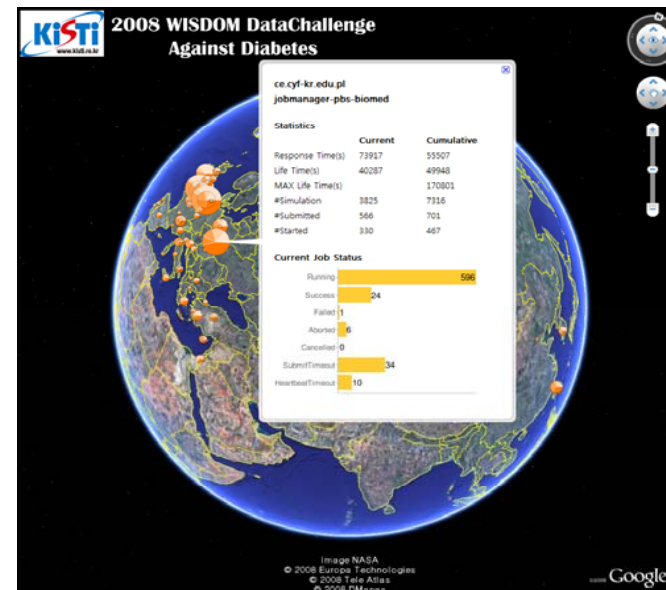
WISDOM Production Environment Monitoring Page

Host: amga.kisti.re.kr | Port: 8823 | InstanceName: DataChallenge

ID	Name	Total CPU	Free CPU	Waiting Jobs	Running Jobs	Response Time	Life Time	#Simulations	#Submitted	#Started	Score
TOTAL		62376	22537	12964	10995			246151	153395	43824	
98	ce.gina.sara.nl jobmanager-pbs-medium	496	43	0	361	002/	50285	9111	1310	1292	9111
17	ce.cyf-kr.edu.pl jobmanager-pbs-biomed	552	19	4	170	27991	49948	9527	1490	1018	9527
86	lgcse02.gridpp.rl.ac.uk jobmanager-lcgpbs-grid500M	2583	1585	11	380	2182	50236	8806	1472	1419	8806
128	ce2.pggrid1.rhul.ac.uk jobmanager-pbs-biomed	400	1	165	0	9361	49480	6352	1541	444	6352
15	gridgate.cs.tcd.ie jobmanager-pbs-eege	691	156	123	437	8611	54533	6082	1427	900	6082
85	ce.ref.man.poznan.pl jobmanager-pbs-biomed	552	258	0	0	24133	37798	5962	1461	349	5962
79	gazon.nikhef.nl jobmanager-pbs-qlong	1152	1	12	385	9710	37822	5757	1443	863	5757
144	serv03.hep.phy.cam.ac.uk jobmanager-lcgpbs-biomed	152	0	6	143	22011	26739	4993	1190	633	4993
80	trekker.nikhef.nl jobmanager-pbs-qlong	1152	0	12	385	13784	42626	4988	1481	768	4988
103	w-ce02.grid.simca.edu.tw jobmanager-lcgpbs-biomed	1480	156	118	100	18367	55944	4791	1476	414	4791
126	ce01.exc.qmul.ac.uk jobmanager-lcgpbs-lcgp_long_x86	1674	801	0	318	3810	50104	4291	1268	1107	4291

Update Period: 60 seconds | Time Remaining: 52 seconds

Job distribution and execution status with CE info



Additional Features relevant to Performance Improvement of WISDOM PE

- **Flexible fault tolerance**
 - Can recover a previous instance using data on AMGA
 - Can increase/decrease the number of agents of a running previous instance
- **Improved Job distribution efficiency**
 - Dynamic management of the list of available RB and CE
 - While the initial job submission keeps going, we can check the status of submitted jobs and resubmit failed jobs
 - Direct collecting the statistical data of sites from WN
 - Its own ranking based on previous statistical data
 - Timeout feature (submission timeout, heartbeat timeout)

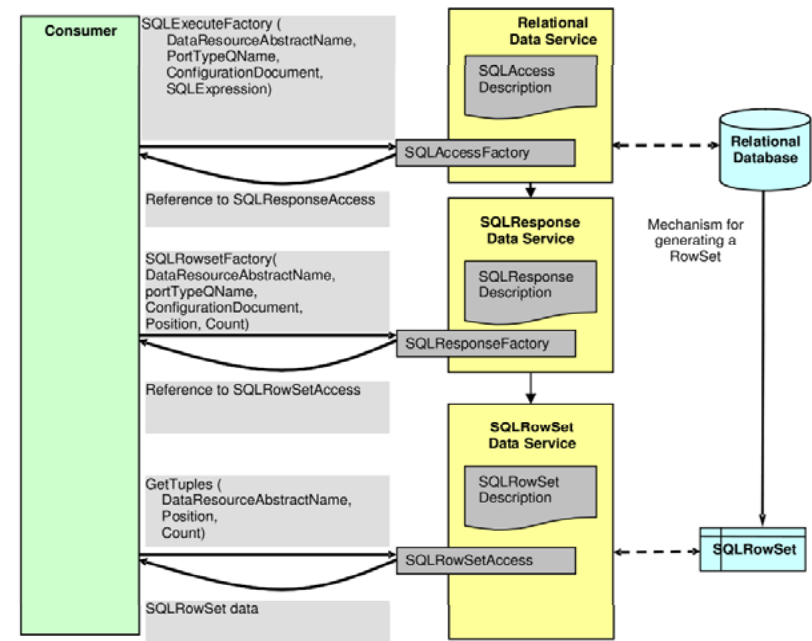
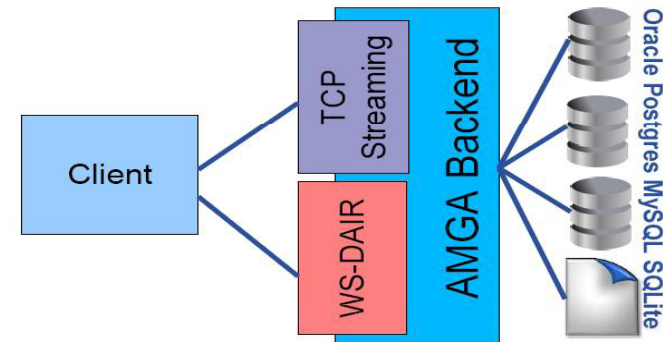
New Developments in AMGA

AMGA 2.0

- **Currently preparing AMGA 2.0 release**
 - To be released later this year
 - Stability of WS-DAIR server needs to be improved before release
- **Feature-complete 1.9 technology preview available**
 - Support for the **import of existing relational tables**
 - **WS-DAIR** frontend
 - **Native SQL** support
 - **Multi-threaded DB backend** with connection pooling

WS-DAIR Support

- WS-DAIR
 - OGF standard for access to relational DB's on the Grid
- Background
 - Enable interoperability to facilitates sharing data
 - AMGA uses its own protocol and message patterns
 - Integration of WS-DAIR in AMGA will make AMGA a relational data source in a WS-based environment!
- AMGA-DAIR Integration
 - Implementation of direct access and indirect access done.
 - Data returned in Sun's WebRowSet specification.
 - All queries use SQL.
 - Clients written in C++ (gSOAP)

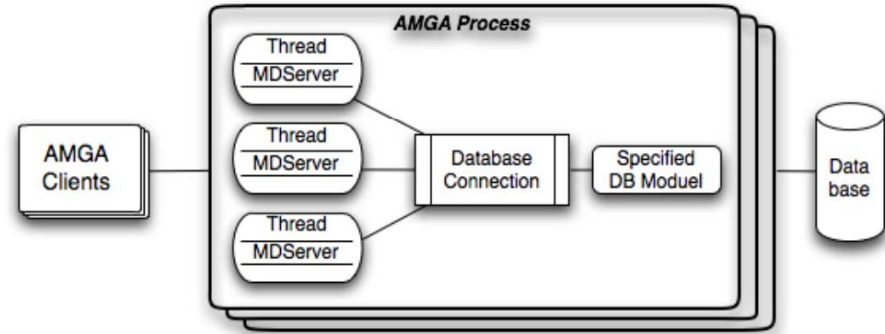
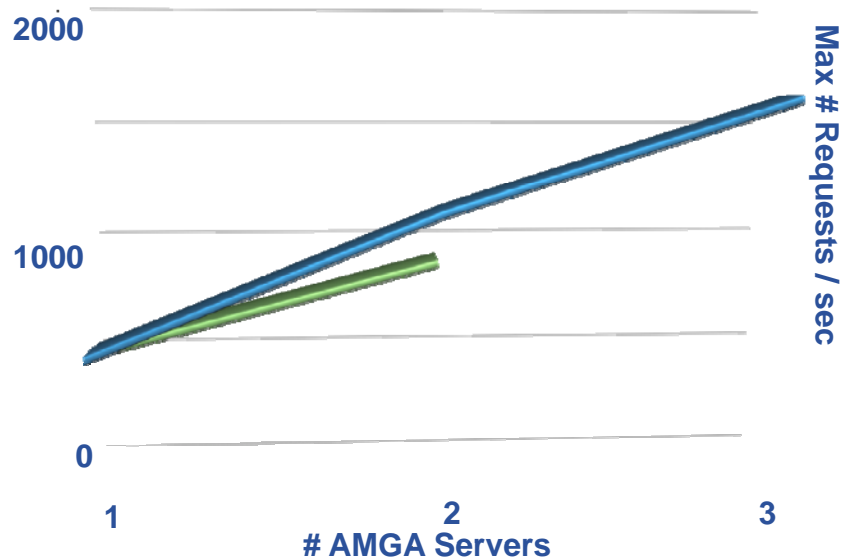


SQL Queries

- **AMGA allows queries now in native SQL**
 - Support for entry level SQL 92 done
 - Some 92 intermediate level supported
 - SELECT, UPDATE, INSERT, DELETE
- **All queries are subject to AMGA access restrictions**
 - All queries are parsed by AMGA, AMGA “understands” security implications
 - Posix ACLs for tables/entries
- **SQL 92 query is translated into backend DB dialect**

Multithreaded & DB Connection Pool Backend Support

- **AMGA 1.3** used one process per client connection
- Processes allow control of misbehaving clients
- But processes cannot share DB connections
 - Limits # concurrent clients



- **AMGA 1.9:** Multiple threads sharing a DB connection encapsulated in multiple processes
- Threads allow better resource usage on DB
- WISDOM measured 1500 Queries/s, more than the required 5million Queries / h peak needed for data challenges

Thank you