Grid Heterogeneity in In-silico Experiments: An Exploration of Drug Screening Using DOCK on Cloud Environments

Wen-wai YIM[a], Shu CHIEN[a], Yasuyuki KUSUMOTO[b], Susumu DATE[c,b], and Jason HAGA[a]

[a] *Dept. of Bioengineering, University of California, San Diego, La Jolla, CA, USA*

[b] *Graduate School of Information Science and Technology, Osaka University, Osaka, Japan*

[c] *Cybermedia Center, Osaka University, Osaka, Japan*

Introduction

- Recent advances in genomic, proteomic, and molecular sciences has led to an explosion of data available. [1]

- Estimated to be growing exponentially, doubling in a six-month period [2]
  - This is faster than the predicted 18-month rate of individual processor doubling (Moore's Law)

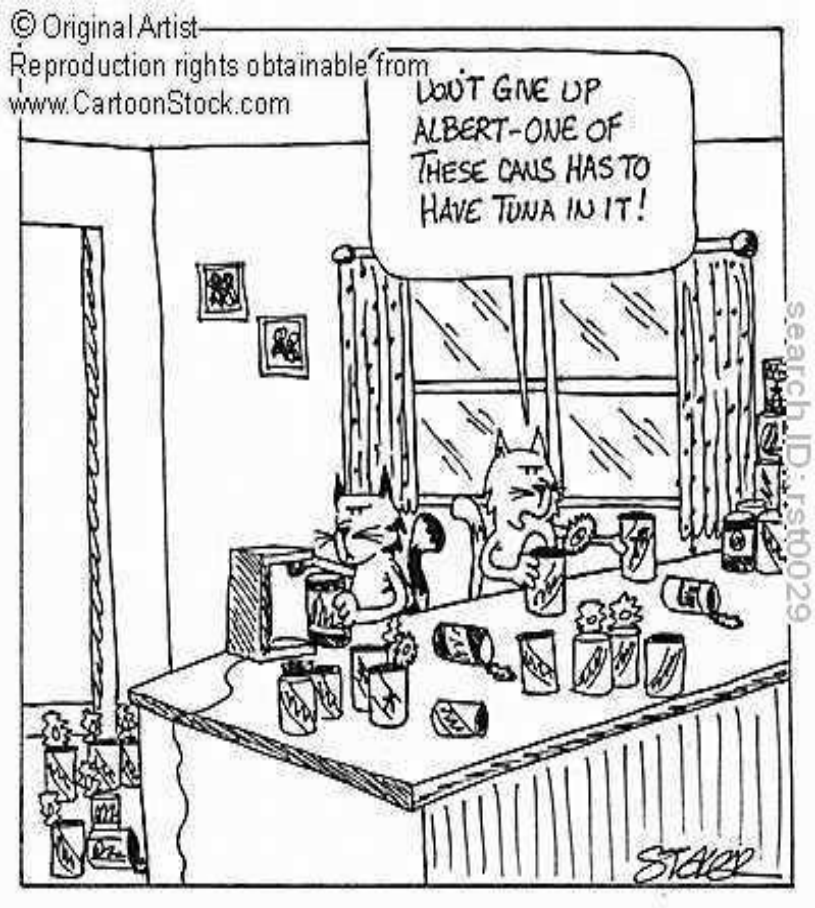
- Drug Discovery: Number of compounds to screen is greater. (~10 million)

Traditional High Throughput Screening (HTS) is no longer a viable option

- Large-scale wet-lab trial and error
- Bleicher, et al on HTS:

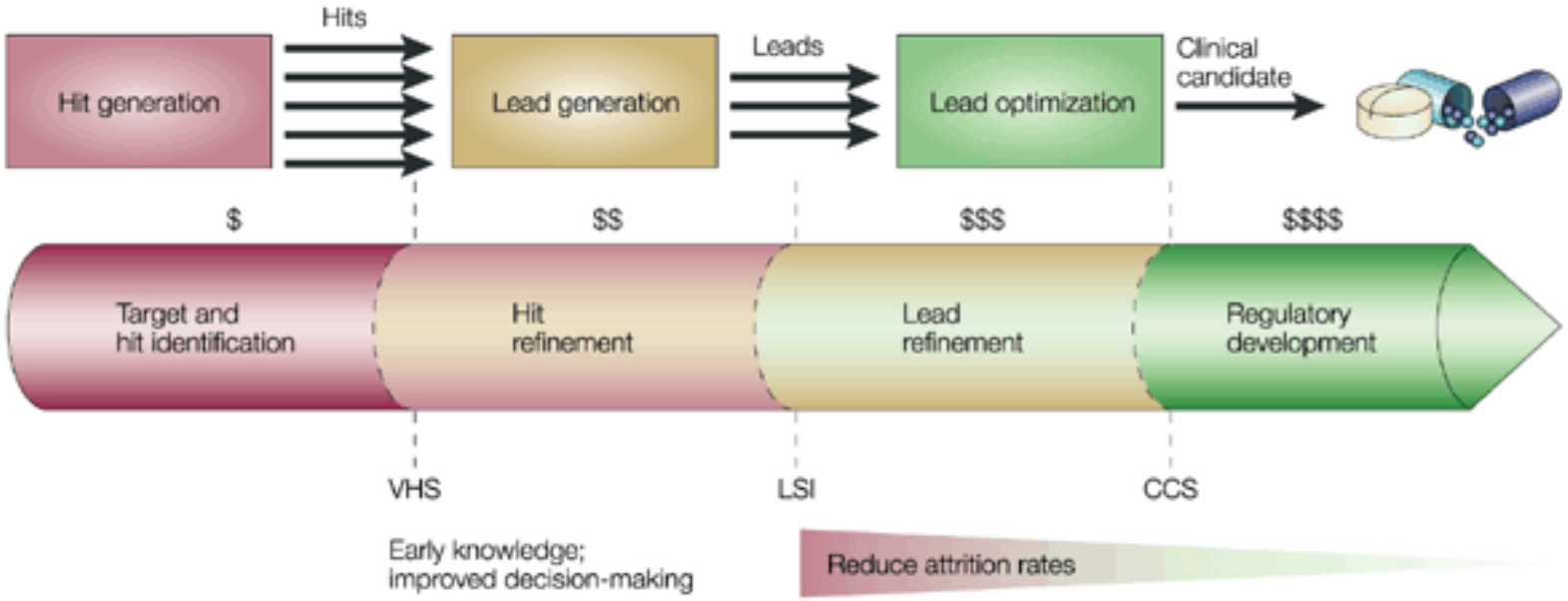
  "Despite massive growth in screening compound numbers over the past 15-20 years, no corresponding increase in successfully launched new chemical entities has resulted" [3]

  - Identified leads do not always meet ADME (absorption, distribution, metabolism, and excretion) requirements
  - Inefficient because screens similar compounds
  - Usually only a few concentrations are used
  - Success threshold is somewhat arbitrary



Virtual Screening

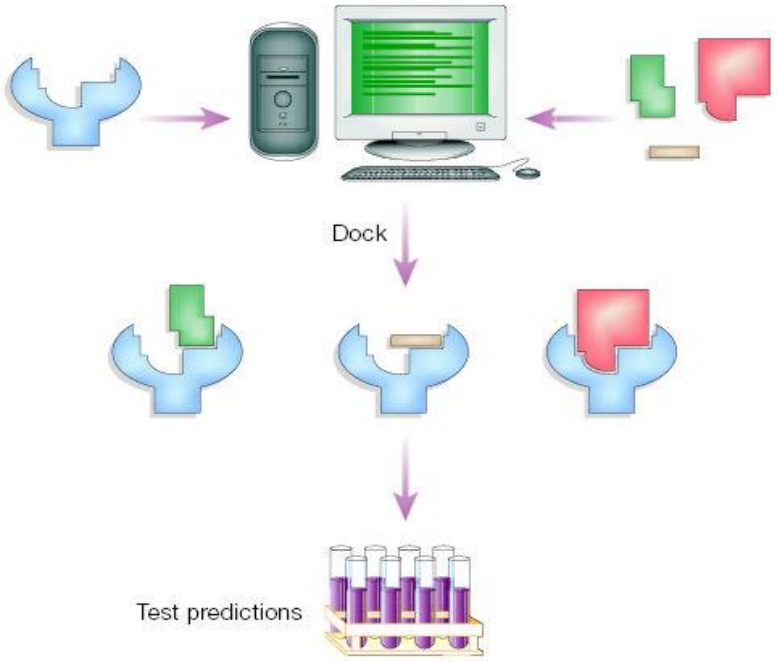
- "In-silico techniques … save an average of $130 million and 0.8 years per drug" [3]
- "Iterative screening can provide 20-times higher enrichment than random screening" [3]



A guide to drug discovery: Hit and lead generation: beyond high-throughput screening
http://www.nature.com/nrd/journal/v2/n5/full/nrd1086.html

Dock6.2, a molecular docking simulation program

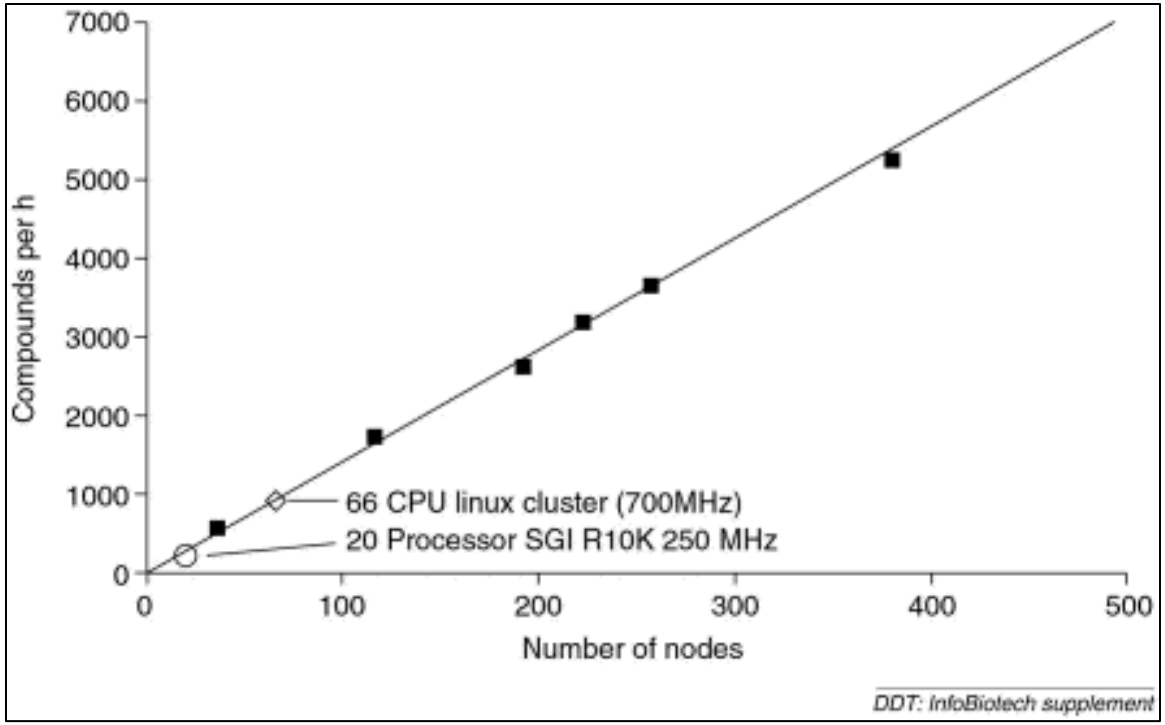
- Developed by UCSF
- Models ligand/receptor interactions
  - Dock score: sum of the van der Waals attractive, van der Waal dispersive and Coulombic electrostatic energies
  - Time
- http://dock.compbio.ucsf.edu/Overview_of_DOCK/index.htm
- Ligands/Receptor from and ZINC and RCSB PDB database



Virtual screening of chemical librariesBrian K. Shoichet Nature 432, 862-865(16 December 2004) doi:10.1038/nature03197back to article

The Grid as a Solution

- Grid technologies provide a cheap solution sharing computational resources.



Grid technologies empowering drug discovery. Drug Discovery TodayVolume 7, Issue 20, 15 October 2002, Pages s176-s180

The Grid Heterogeneity Problem

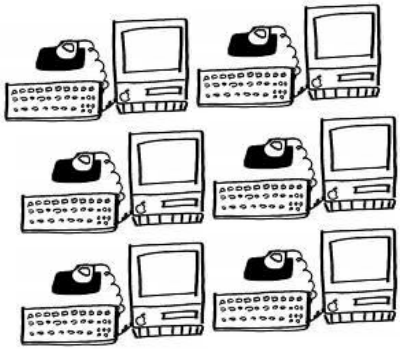
- Like in cooking:: not everyone has exactly the same brand or type of ingredients in their own homes

  - >> RESULT
    - SLIGHTLY DIFFERENT OUTCOMES

For computer clusters on the Grid, this is also true… [current system]

DOCK RESULTS FOR THREE DIFFERENT CLUSTERS

Cluster 1



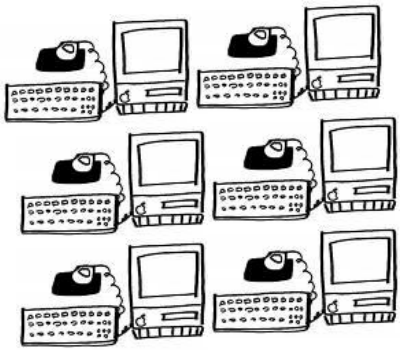
Cluster 2

Cluster 3

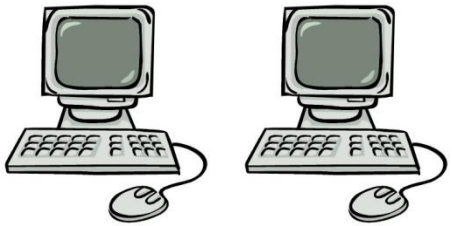
	cluser 1	cluster 2	cluster 3
ZINC00310164	-85.142899	-53.300598	-91.583946
ZINC02353150	-47.200092	-44.72197	-42.266335
ZINC01786999	-40.391968	-35.641331	-37.995266
ZINC01060696	-32.584995	-37.99453	-38.034916

Implications of heterogeneous performance in terms of a Grid environment. [current system]

Cluster 2



Cluster 3



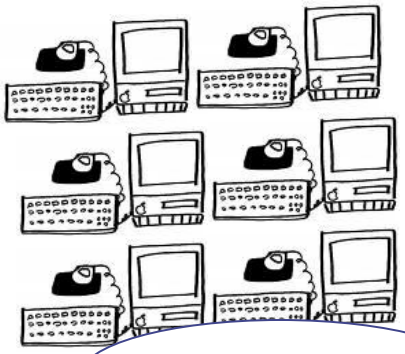
Cluster 1

ZINC00310164
ZINC01060696
ZINC01786999
ZINC02353150

ZINC00310164
ZINC01060696
ZINC01786699

Implications of heterogeneous performance in terms of a Grid environment. [current system]

Cluster 2



ZINC01060696

ZINC0178699

ZINC02353150

Cluster 3

ZINC00310164

- Cluster 1

Docking scores from distributed Grid job. [current system]

	cluser 1	cluster 2	cluster 3	GRID collaboration
ZINC00310164	-85.142899	-53.300598	-91.583946	-85.142899
ZINC02353150	-47.200092	-44.72197	-42.266335	-44.72197
ZINC01786999	-40.391968	-35.641331	-37.995266	-35.641331
ZINC01060696	-32.584995	-37.99453	-38.034916	-38.034916

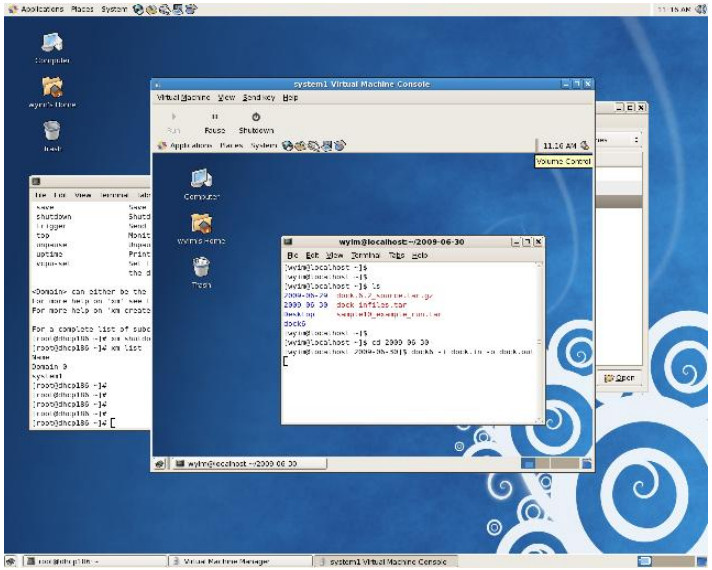
Ligand rankings from distributed Grid job.

cluster 1	cluster 2	cluster 3	GRID collaboration
ZINC00310164	ZINC00310164	ZINC00310164	ZINC00310164
ZINC02353150	ZINC02353150	ZINC02353150	ZINC02353150
ZINC01786999	ZINC01060696	ZINC01060696	ZINC01060696
ZINC01060696	ZINC01786999	ZINC01786999	ZINC01786999

Contributing factors to heterogeneity

- (1)Random Number Generator works differently for different compilers
- (2)Roundoff in calculations due to different libraries used during compilation
- (3)Precision roundoff due to computer architecture

How does a virtual machine help?



A virtual machine is like running a self-contained OS on a physical machine.

- Provides a layer of abstraction that makes it run, for the most part, independent of the host machine.

- As a self-contained OS image, a virtual machine, complete with all its configured functions can be deployed on many host computers.

Experiment Overview

Purpose

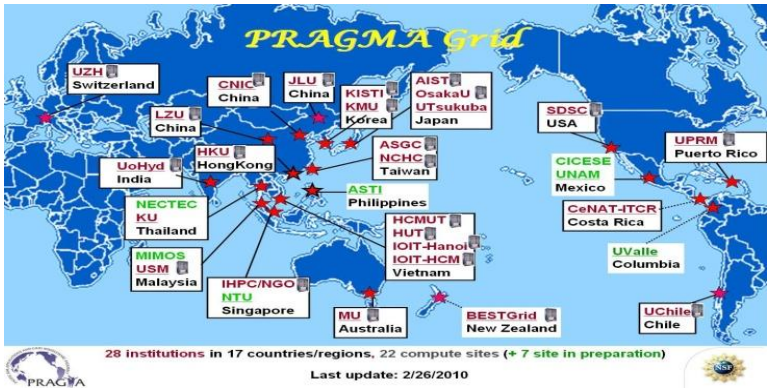
To quantify virtual screening (VS) results and performance on (1) a grid and (2) on virtual clusters in a grid environment.

Methods

1. Run DOCK6.2 on a operational grid organization to quantify DOCK6.2 molecular screening differences.
2. Run DOCK6.2 virtual screen on two different physical clusters.
3. Run DOCK6.2 VS on virtual clusters (VC) installed on top of the physical clusters.
4. Compare performance of DOCK6.2 on the physical cluster and VC.

Running DOCK6.2 on an operational Grid:

The PRAGMA Grid

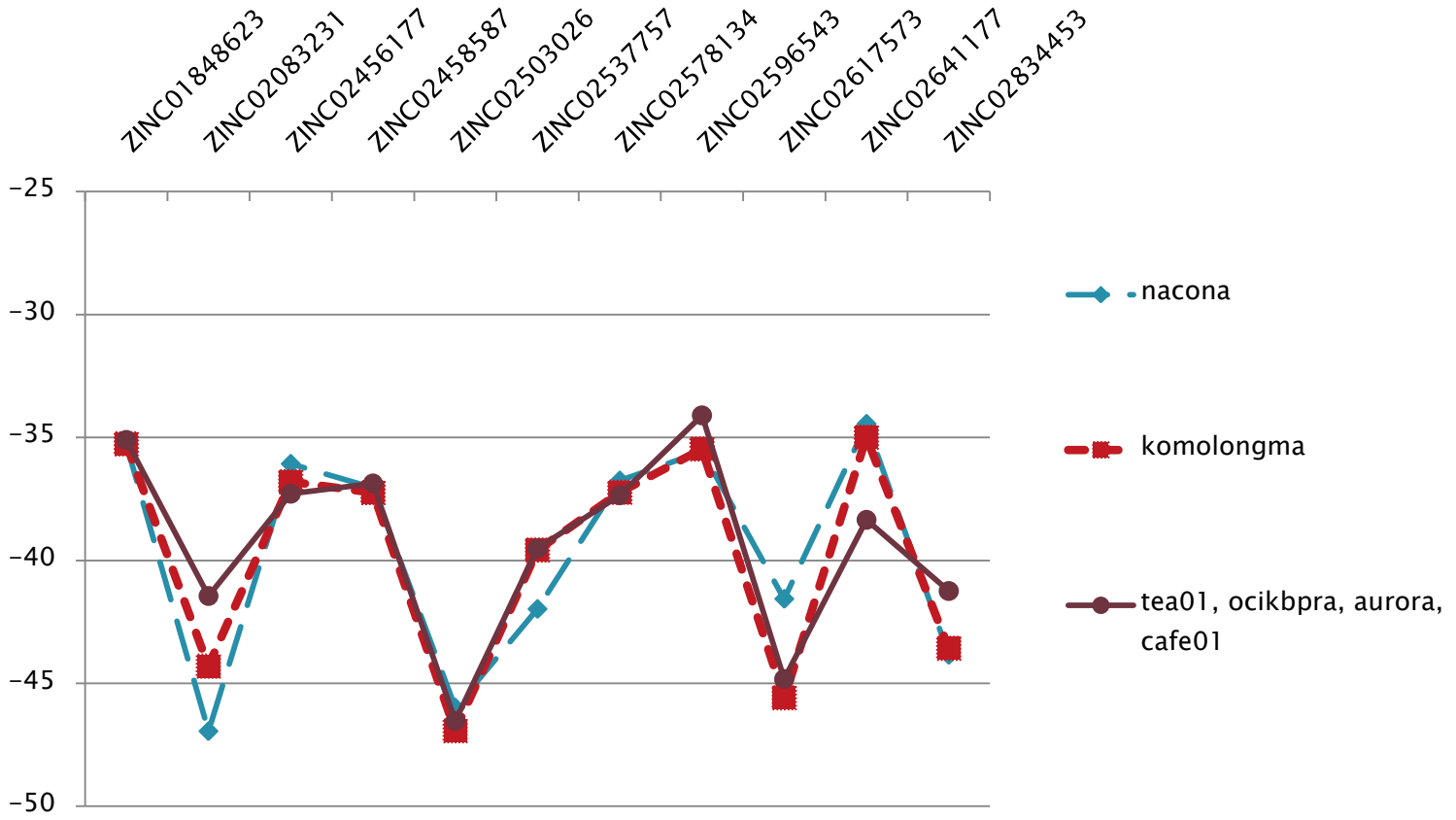


Institution Name	Country /Region	Host Name	Total Nodes	Total CPUs	Memory per Node (GB)	CPU Model	CPU Speed (MHz)
NCHC	Taiwan	nacona00	9	18	3.8	x86_64	3200
OSAKAU	Japan	tea01	40	80	1.0	i686	1400
OSAKAU	Japan	cafe01	20	40	2.0	i686	2800
UPRM	Puerto Rico	komolongma	60	110	1.0	i686	1200
USM	Malaysia	aurora	17	34	1.0	i686	1396
UZH	Switzerland	ocikbpra	10	20	2.0	i686	2800

Results I. PRAGMA Grid

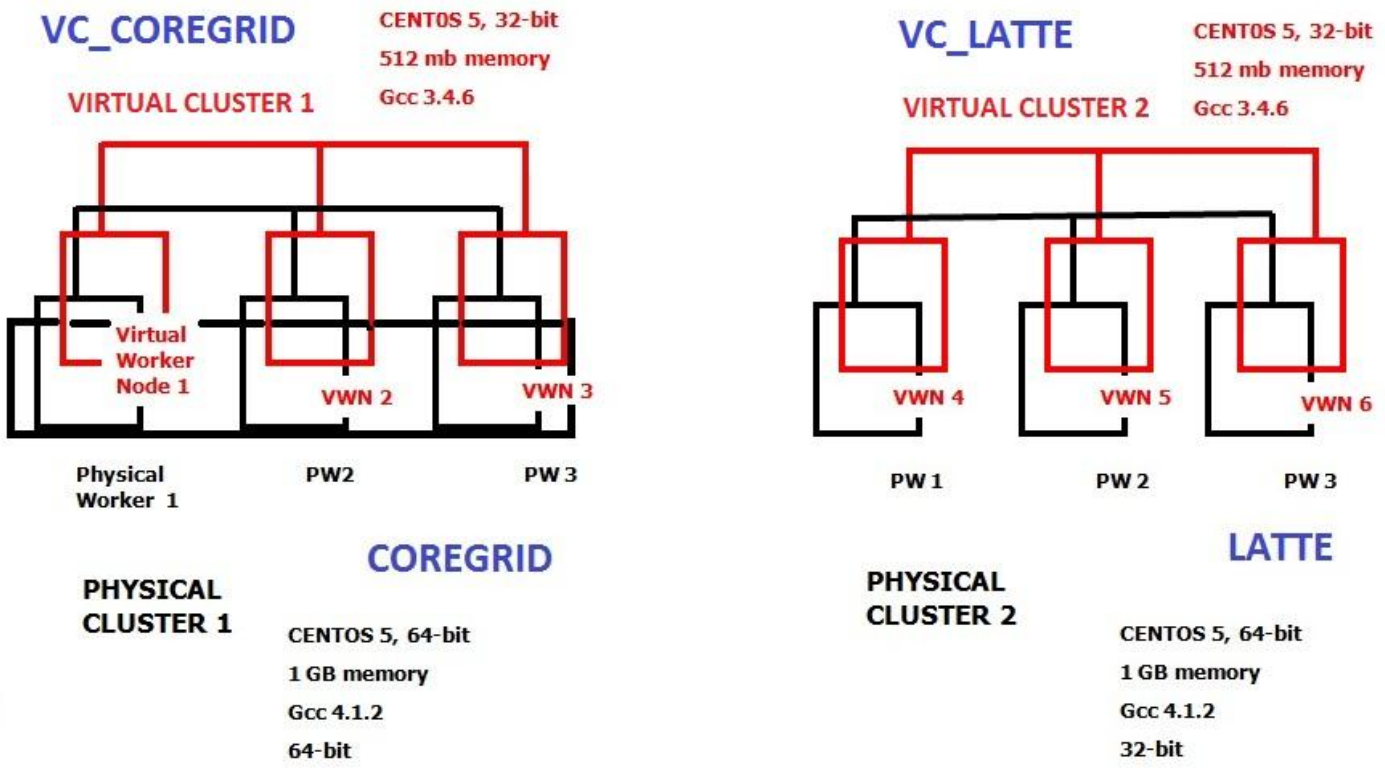
	developer	***nacona***	***komolong.***	***ocikbpra***	***aurora***	***cafe01***	***tea01***
OS-bit		64	32	32	32	32	32
machine bit		64	32	32	32	32	32
gcc version		3.4.6	4.1.2	3.4.6	3.4.6	3.4.6	3.4.6
ZINC ID							
00013564	**10.80194**	-9.40022	10.80194	-9.40021	-9.40021	-9.40021	-9.40021
00150863	**21.13914**	-12.46722	21.13914	21.13914	21.13914	21.13914	21.13914
00152265	**30.23836**	30.23803	30.23836	19.05393	19.05393	19.05393	19.05393
00157111	**-12.9166**	-12.9166	-12.9166	-15.5221	-15.5221	-15.5221	-15.5221
00157152	**-10.1374**	-10.1374	-10.1374	-10.1374	-10.1374	-10.1374	-10.1374
00157402	**168513.6**	168506.4	168513.6	145519.8	145519.8	145519.8	145519.8
00157467	**-8.70667**	-8.70678	-8.70667	-8.70667	-8.70667	-8.70667	-8.70667
00157960	**847.3721**	847.3558	847.3721	847.3721	847.3721	847.3721	847.3721
00158442	**52.56588**	52.56579	52.56588	52.56588	52.56588	52.56588	52.56588
00158751	**-1.05811**	-1.05816	-1.05811	-1.05811	-1.05811	-1.05811	-1.05811
01555236	**503.250**	501606080	503.250				

Results I. (continued)



For 100 compounds the average discrepancy was 1.0 and average percent variation was 2%.

Running DOCK6.2 on virtual clusters: A Local Grid



Virtual Machines Created From Xen Hypervisor

Results II. Virtual Clusters

	developer	*VC_coregrid*	*VC_latte*	*coregrid*	*latte*
OS-bit		32	32	64	32
machine bit		64	32	64	32
gcc version		3.4.6	3.4.6	4.1.2	4.1.2
ZINC ID					
00013564	**10.801939**	-9.400211	-9.400211	-9.400218	10.801939
00150863	**21.139143**	21.139143	21.139143	-12.467216	21.139143
00152265	**30.238361**	19.05393	19.05393	30.238028	30.238361
00157111	**-12.916615**	-15.522113	-15.522113	-12.916644	-12.916615
00157152	**-10.137384**	-10.137392	-10.137392	-10.137392	-10.137384
00157402	**168513.625**	145519.7813	145519.7813	168506.4063	168513.625
00157467	**-8.706671**	-8.706671	-8.706671	-8.706783	-8.706671
00157960	**847.37207**	847.372131	847.372131	847.355774	847.37207
00158442	**52.56588**	52.565876	52.565876	52.565788	52.56588
00158751	**-1.058107**	-1.058107	-1.058107	-1.058161	-1.058107
01555236	**503.249725**			501606080	503.249725

Results II. (Continued)

	coregrid	latte	VC_coregrid	VC_latte
DOCK Time Per Molecule	13.5 ± 0.4	48 ± 0.4	18.8 ± 0.3	52 ± 7
Total Execution Time	674.6 ± 0.5	2420 ± 5	953 ± 2	2460 ± 6

Conclusion

- Virtual clusters can be a good solution to the grid heterogeneity problem.
- Virtual cluster performance/overhead costs in virtual screening are acceptable.
- Not only would virtual clusters enforce conformity in performance among different platforms, it can provide a level of autonomy and independence to users without compromising the security and preferences of other users.

Acknowledgements

University of California, San Diego

Dr. Gabriele Wienhausen

Dr. Jason Haga

Marshall Levesque

Dr. Peter Arzberger

Teri Simas

Osaka University

Dr. Shinji Shimojo

Dr. Susumu Date

Yasuyuki Kusumoto

Dr. Kohei Ichikawa

Funding Provided the National Science Foundation (IOSE-0710726)

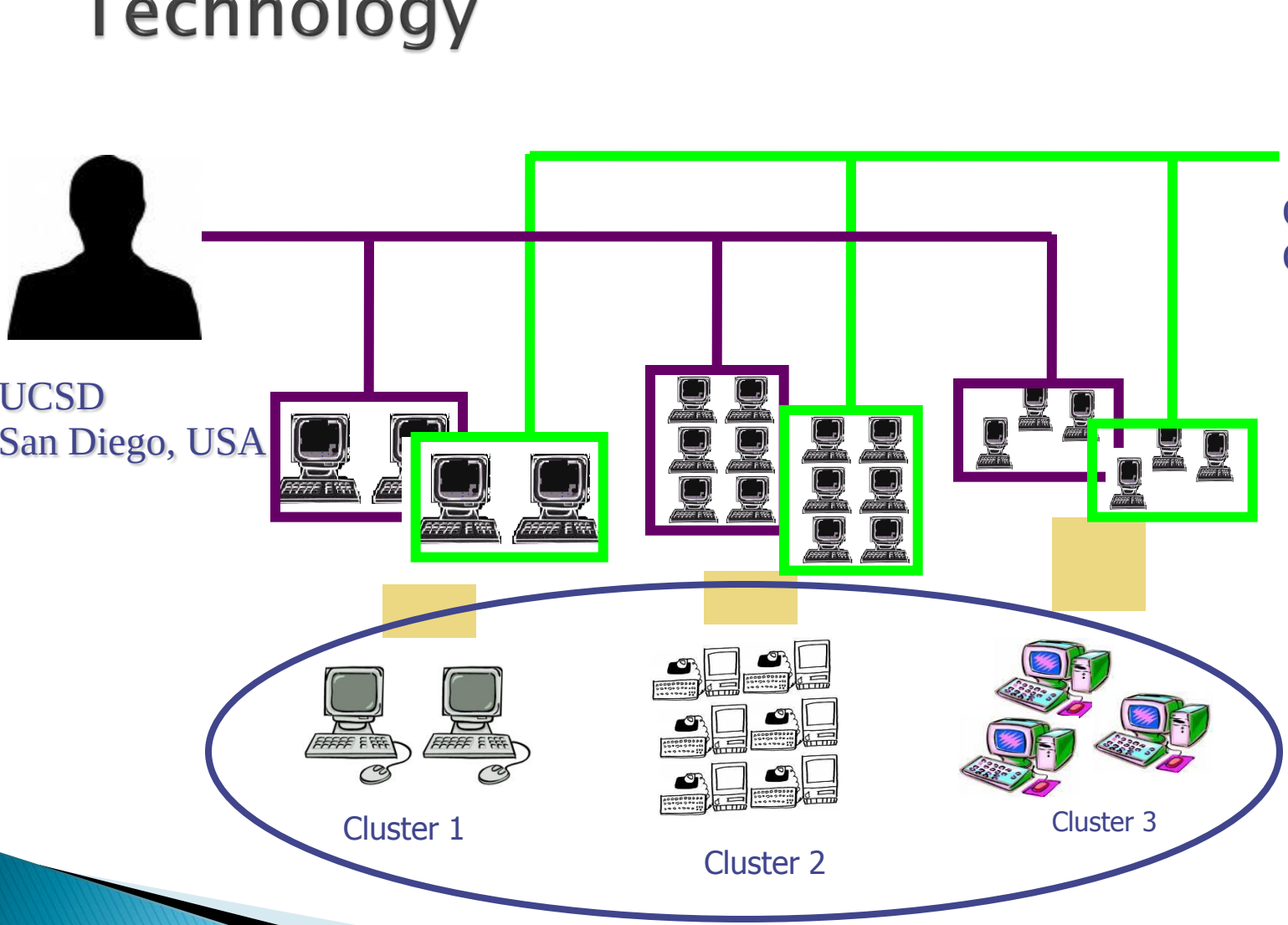
Thank you for your attention.

Any questions, please … ?

Future Challenges

- Stable cross-grid network solution.
- Resource sharing scheduler for cluster spawning.
- Automation for the creation and deploying of virtual cluster on clusters.
- Larger cooperation on the part of on grid organizations to experiment in virtual cluster technology and identify the exact overhead cost and feasibility of virtual cluster modality of resource sharing.

Goal of Virtual Cluster Overlay Technology



Grid Resources: Shared

References

[1] A Krishnan, A Survey of Life Sciences Applications on the Grid. New Generation Computing, 2004: p. 111-126

[2] A Chien, I.F.a.D.G., Grid technologies empowering drug discovery. Drug Discovery. Today, 2002: p. 176-180.

[3] KH Bleicher, H.J.B., K. Muller, A.l. Alanine, Hit and lead generation: beyond high-throughput screening. . Nat Rev Drug Discov, 2003. 2: p. 369-378.

[4] BK Shoichet. Virtual screening of chemical libraries. Nature 432, 862-865(16 December 2004) doi:10.1038/nature03197back to article

[5] A. Chien, I.F.a.D.G . Grid technologies empowering drug discovery. Drug Discovery Today. Volume 7, Issue 20, 15 October 2002, Pages s176-s180