The Erasmus Computing Grid

Building a Super-Computer

for *Free*

Dr. Tobias A. Knoch & Luc V. de Zeeuw

with

R. de Graaf, A. Abuseiris, W. Burger, R. Ali, & P. den Brock

Biophysical Genomics & GLOBE Consortium, Cell Biology & Genetics, Erasmus MC

Hogeschool Rotterdam

ECG Basic Grid Structure

The ECG consists of the computer pools of the Erasmus MC and the Hogeschool Rotterdam. The client computers are controlled by the general ECG Emperor of the grid at EMC and a local master at HR. "Condor" is used as middleware since it is very well established and open-source.

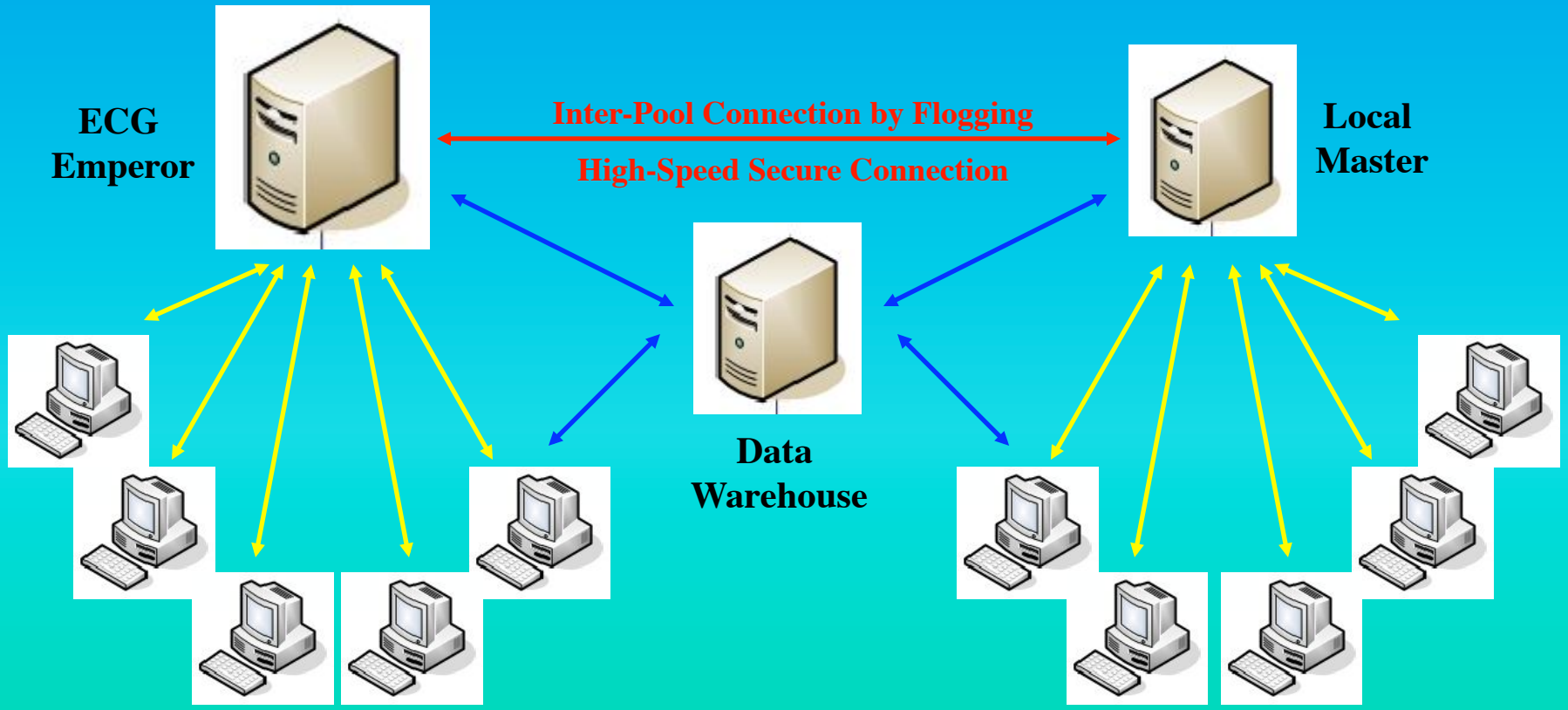
Erasmus Computing Grid

Erasmus MC
University Medical Center Rotterdam

HOGESCHOOL ROTTERDAM

~9000 CPUs or ~18000 vHosts

~3500 CPUs or ~7000 vHosts



ECG Basic Grid Structure

Erasmus Computing Grid

The ECG consists of the computer pools of the Erasmus MC and the H
client computers are controlled by the general ECG Emperor of
master at HR. "Condor" is used as middleware since it is ver

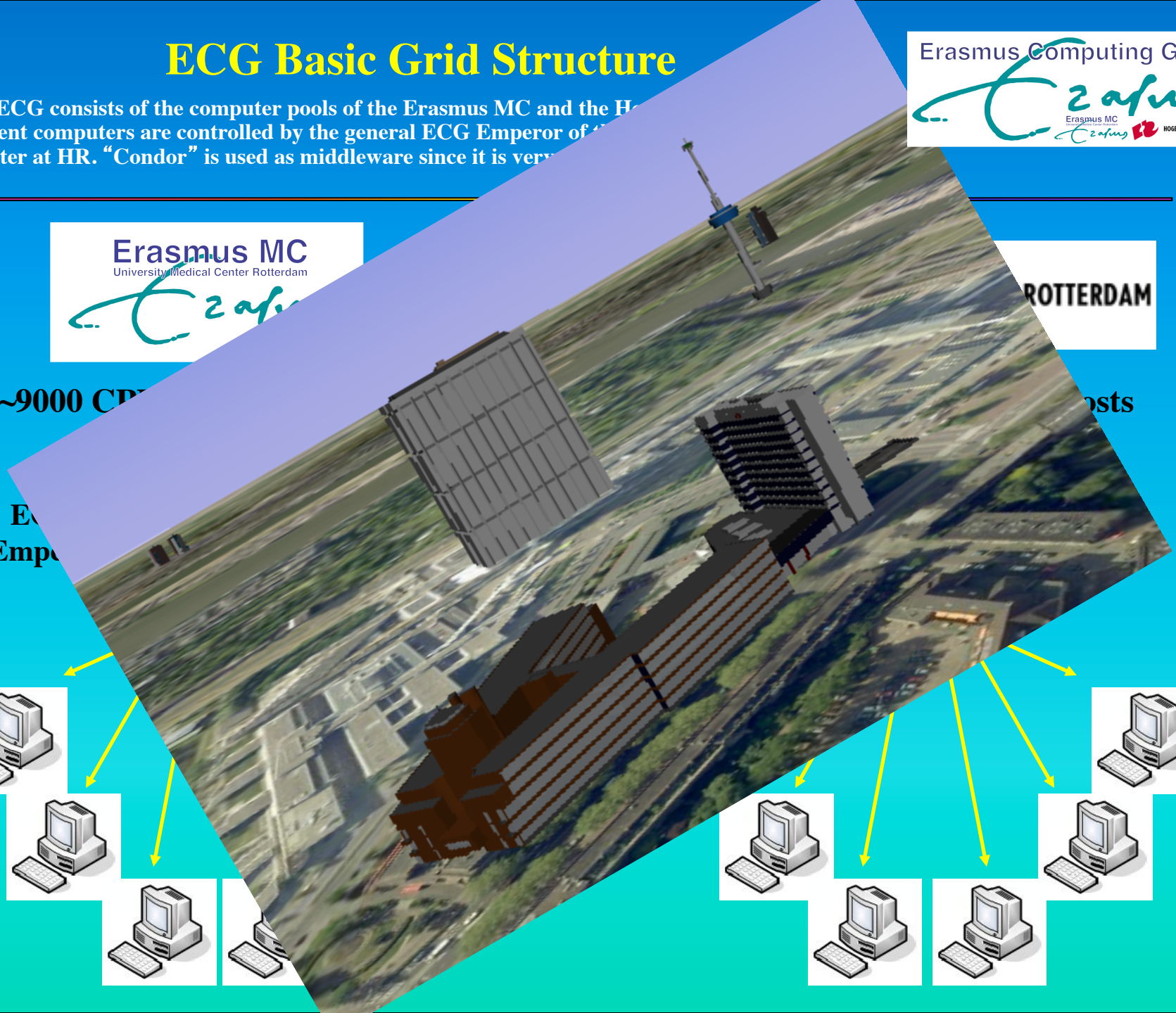
Erasmus MC
University Medical Center Rotterdam

ROTTERDAM

~9000 CP

osts

E
Emp



World-Map of Condor Grids

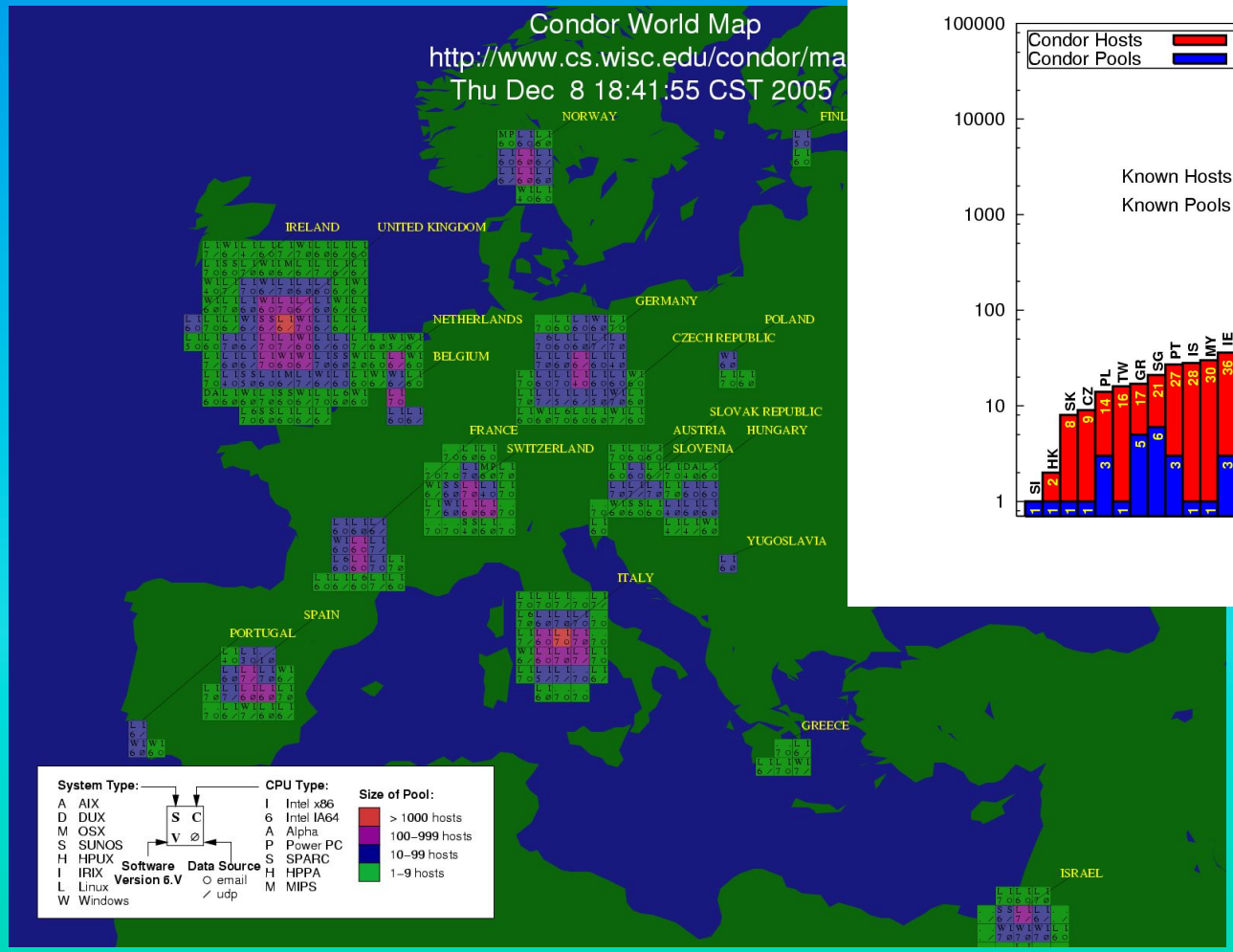
Condor is used as the standard for batch systems on many real parallel super-computers as well as in many distributed trivial parallel computing grids. It is running in very diverse environments of the highest security levels as e.g. governmental or industry R&D.

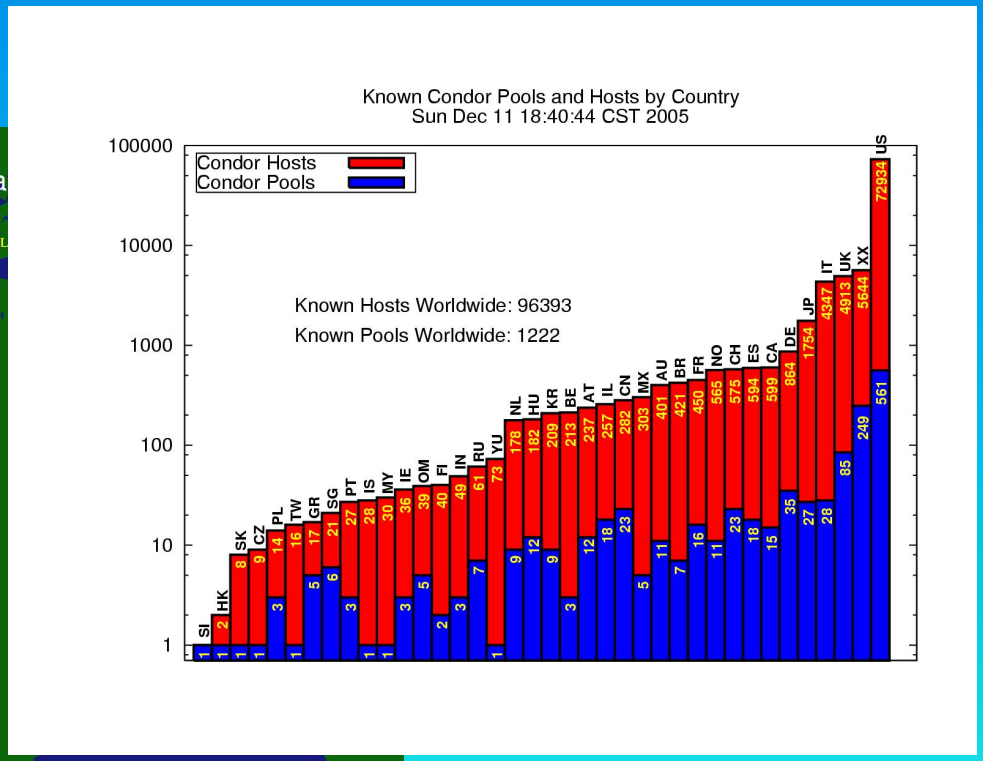
With a potential capacity of currently ~20 to 25 Tera flops the ECG is one of the biggest grids!

Erasmus Computing Grid

Erasmus MC

HOGESCHOOL ROTTERDAM





Based on pool operator reports only!

ECG Projects and Users

The projects and users currently run on the ECG are mainly basic research and application development in the biological and medical areas. Currently there are ~15 projects and ~10 users of the ECG with large expansion already in sight.

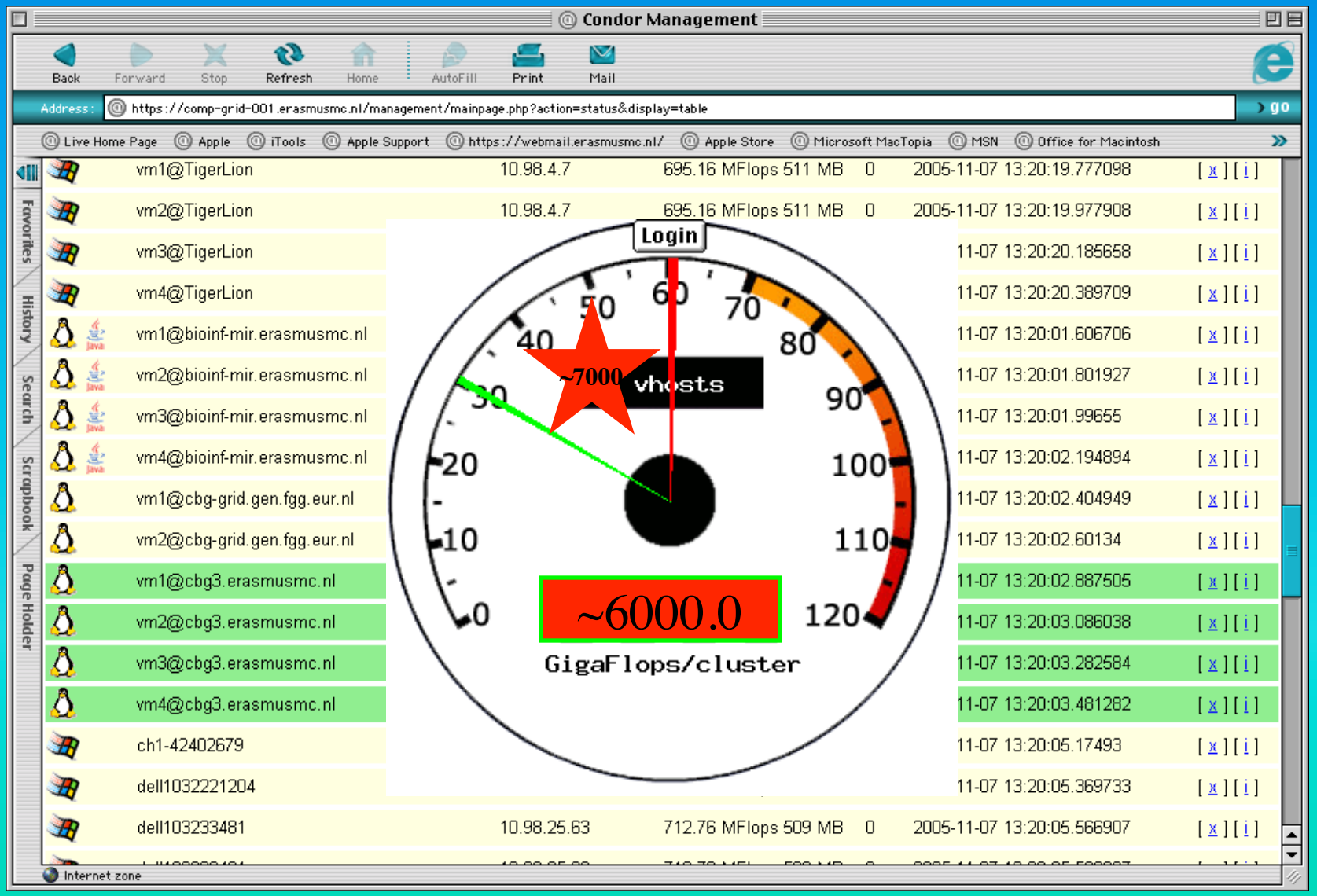
The ECG is one of the few grids on this level of complexity !

Image Analysis	->	J. v. d. Lei and W. Nissen, Dept. Medical Informatics
DNA Sequence Correlation	->	T. A. Knoch, Dept. Cell Biology & Genetics
Genome Duplication Analysis	->	T. A. Knoch, Bert Eussen, Dept. Cell Biology & Genetics
Epidemiology studies	->	Dept. Public Healt
CT Image Analysis	->	H. v. Leuwen, Dept. Internal Medicine
Structural Genome Simulations	->	T. A. Knoch, Dept. Cell Biology & Genetics
Protein Motive Patterns	->	G. Jenster, Dept. Urology
Mass Spektrometre Analysis	->	P. Verijzer, Center for Biomics
Chip & Sequencing Analysis	->	W. van Ijcken, Center for Genetics
Expression Array Analysis	->	A. Uiterlinden, M. Moorhouse
Protein Structure Prediction	->	C. Leslie, Dept. Cell Biology & Genetics
Diffusion of Molecules	->	A. Houtsmueller, Dept. Pathology
Microscopic Image Analysis	->	G. v. Capelle, Dept. Cell Biology

ECG - Middleware

To maintain the ECG data bases for hosts, users and projects were created. They allow to track e.g. the status of the local clients and gives global statistical data about performance. A dedicated job-submission and accounting system is currently implemented. The ECG is also working together with the EGEE, Big-Grid and is a partner of the German nationwide MediGRID / D-Grid.

Erasmus Computing Grid



Security of the ECG

One of the major issues concerning grids is to create a security level sufficient for all involved parties as we did for the ECG. The technology, culture and even the psychology of security have been and are treated very seriously by the ECG and are to the highest achievable level guarantying the integrity and privacy of the system and data !

Erasmus Computing Grid

Erasmus MC

HOGESCHOOL ROTTERDAM

Security Technology

- system hardening
- firewalling
- encrypted network communications
- intrusion detection monitoring
- logging of session
- auditing and testing of applications
- virus/trojan checking of in-/output
- authentication

Security Culture

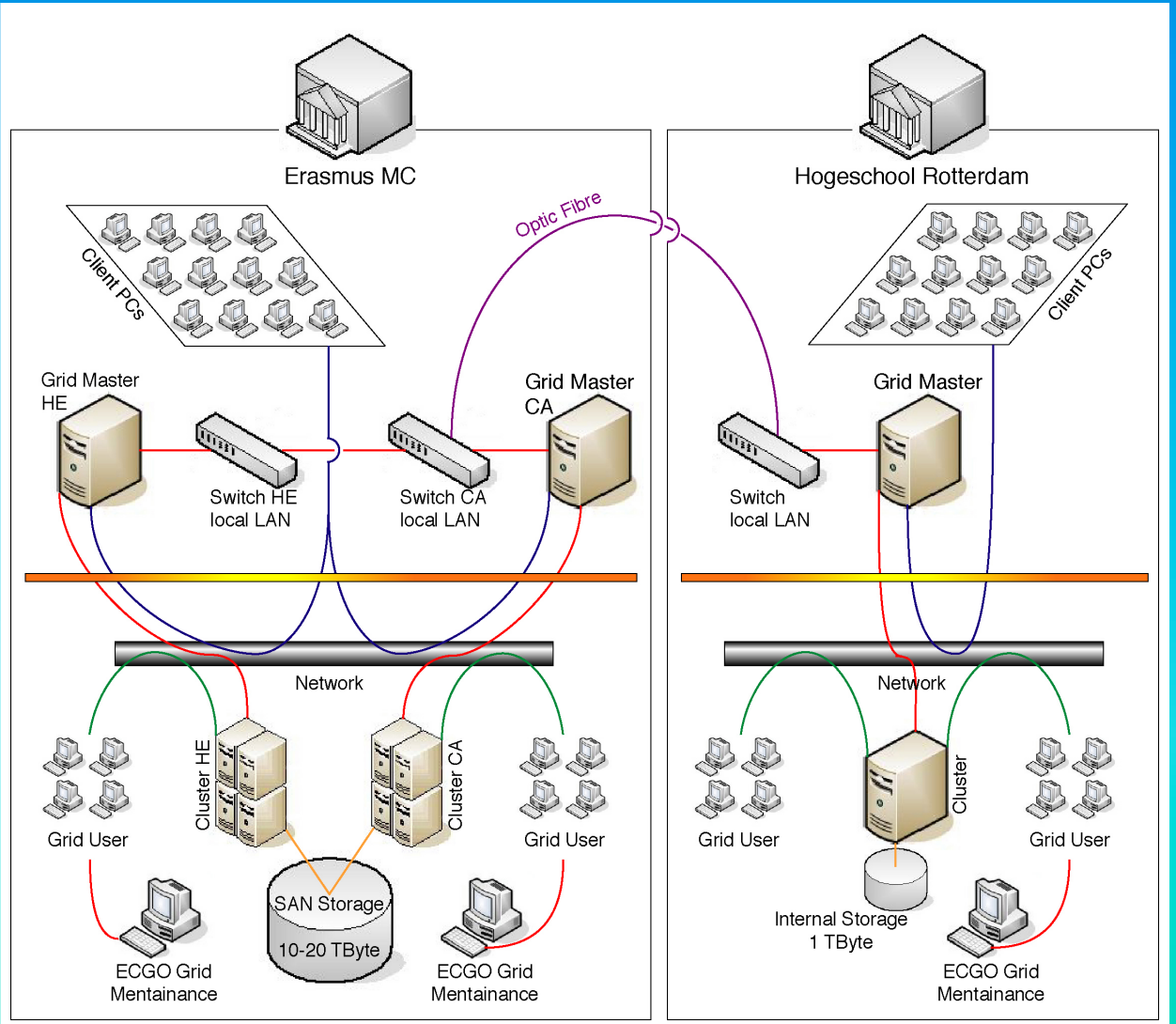
- trained employees
- secure programming
- change management system
- constant check of procedure
- constant test and training of employees
- transparent management structure
- transparent infra structure
- NAN and ISO certification

The highest security level guarantying privacy is just good enough for us.
&
We still would like to have it better.

ECG - Infrastructure

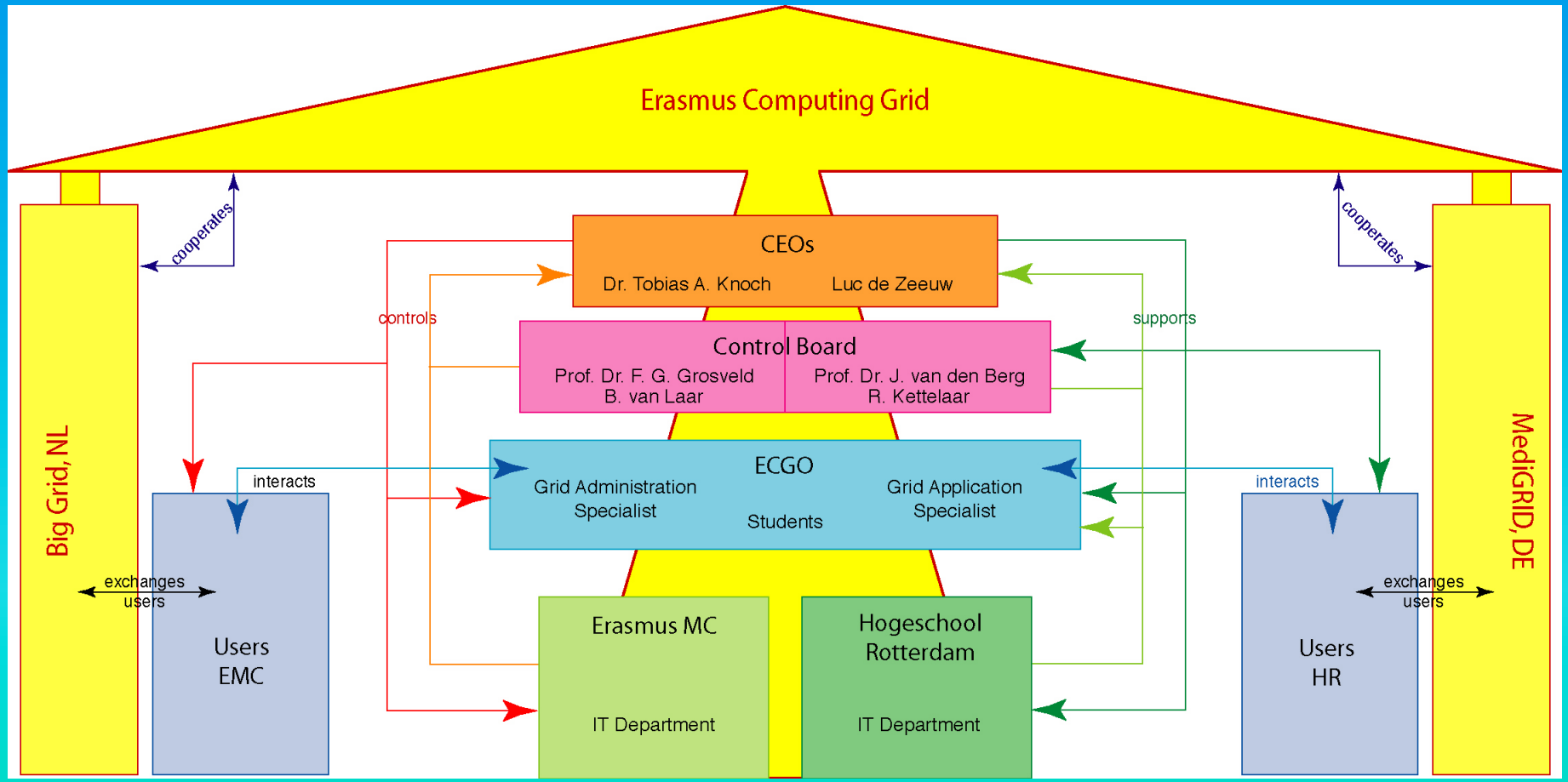
The new infrastructure of the ECG will consist of a dedicated architecture respecting client, user and management security and efficiency in respect to the fact that the ECG is one of the largest grids in the world with the highest possible degree of complexity.

The new ECG infrastructure guaranties the exploitation of the capabilities to the maximum !



ECG - Management

To guaranty the integrity and efficiency of the ECG, a dedicated management structure has been put in place with corresponding checks and balances. The organization is made such, that also collaborations with other grids as e.g. the Dutch Big-Grid or the German nationwide MediGRID / D-Grid can be exploited most efficiently.



ECG - Business Concepts

For the future the ECG has four major business concepts: the research, education, diagnostics and industry arenas. For all of these, separate business concepts, work-flows and user managements are necessary and need to be integrated to allow maximum efficiency.

The ECG is one of the few grids on this level of complexity !

Research:

- genomic and proteomic analysis
- epidemiology
- image analysis, e.g. Applied Molecular Imaging (AMI)

Education:

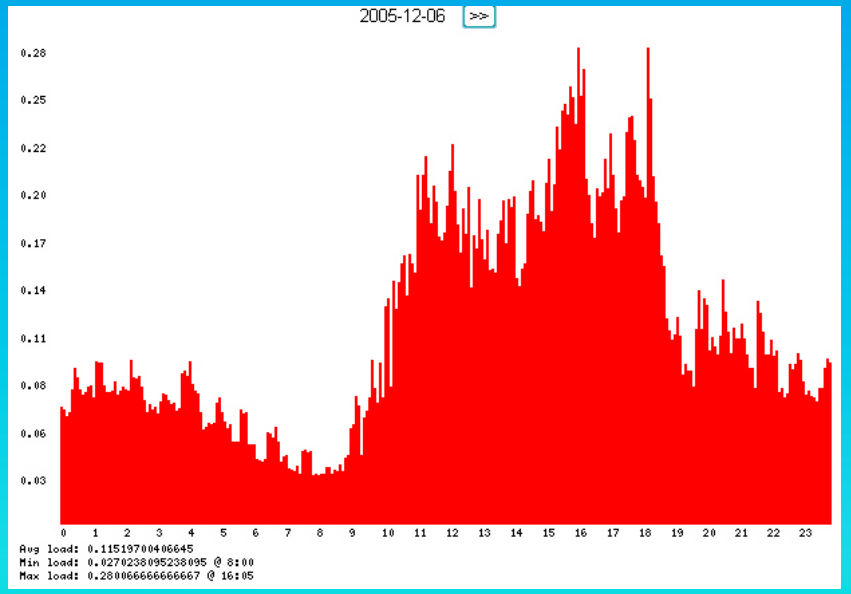
- training of the coming grid generation of IT specialists
- developing new concepts for grid computing

Diagnostics:

- clinical image and data analysis
- operation planning and operation support

Industry:

- brokerage of computing resources



ECG - Budget

The ECG paradigm is the most effective return of investment for computational efforts existing, since the client infrastructure is already in place. Also in respect to the Kyoto protocol the ECG is more efficient due to the initial hardware production and overall energy neutrality during production.

A comparable super-computer would cost in the order of 20-30 Million Euro.!

	2007 [k€]	2008 [k€]	2009 [k€]	Total [k€]
Personal:				
Senior Researcher	110.0	110.0	110.0	330.0
Grid Administrator Specialist	78.0	78.0	78.0	234.0
Grid Application Specialist	78.0	78.0	78.0	234.0
Sub-Total Personal:	*266.0*	*266.0*	*266.0*	*798.0*
ECG-Office:				
ECGO Space	2.0	2.0	2.0	6.0
ECGO Equipment	3.0	3.0	3.0	9.0
ECGO Travel and Marketing	7.0	7.0	7.0	21.0
Sub-Total ECGO:	*12.0*	*12.0*	*12.0*	*36.0*
ECG-Central Infrastructure:				
Master Servers (2; redundant)	12.0	0.0	0.0	12.0
Connecting Switches (2; redundant)	12.0	0.0	0.0	12.0
Fiber Optic Connection to HR	5.0	2.5	2.5	10.0
User Cluster Server (6; redundant)	30.0	0.0	0.0	30.0
Working SUN Storage	40.0	40.0	0.0	80.0
Sub-Total ECG-Central Infrastructure:	*99.0*	*42.5*	*2.5*	*142.0*
Total	**366.0**	**320.5**	**280.0**	**978.0**

ECG - Budget

The ECG paradigm is the most effective return of investment for computational efforts existing, since the client infrastructure is already in place. Also in respect to the Kyoto protocol the ECG is more efficient due to the initial hardware production and overall energy neutrality during production.

A comparable super-computer would cost in the order of 20-30 Million Euro.!

	2007 [k€]	2008 [k€]	2009 [k€]	Total [k€]
Personal:				
Senior Researcher	110.0	110.0	110.0	330.0
Grid Administrator Specialist	78.0	78.0	78.0	234.0
Grid Application Specialist	78.0	78.0	78.0	234.0
Sub-Total Personal:	*266.0*	*266.0*		
ECG-Office:				
ECGO Space	2.0	2.0		
ECGO Equipment	3.0	3.0		
ECGO Travel and Marketing	7.0	7.0		
Sub-Total ECGO:	*12.0*	*12.0*		
ECG-Central Infrastructure:				
Master Servers (2; redundant)	12.0	0.0		
Connecting Switches (2; redundant)	12.0	0.0		
Fiber Optic Connection to HR	5.0	2.5		
User Cluster Server (6; redundant)	30.0	0.0		
Working SUN Storage	40.0	40.0		
Sub-Total ECG-Central Infrastructure:	*99.0*	*42.5*		
Total	**366.0**	**320.5**	**280.0**	**978.0**

ECG-Central Infrastructure COH-Proposal:				
Master Servers (2; redundant)	12.0	12.0	12.0	36.0
Connecting Switches (2; redundant)	12.0	0.0	0.0	12.0
Fiber Optic Connection to HR	5.0	2.5	2.5	10.0
User Cluster Server (6; redundant)	30.0	30.0	30.0	90.0
Working SUN Storage	80.0	80.0	80.0	240.0
Sub-Total ECG-Central Infrastructure COH-Proposal:	*139.0*	*124.5*	*124.5*	*388.0*
Total COH-Proposal	**406.0**	**402.5**	**402.5**	**1,210.0**
Difference to Our Proposal: Assuming Free Storage for Research:				**∂ = +232.0** **∂ = +312.0**

Acknowledgements

Workforce:

Rob de Graaf (ECG administration)
Anis Abuseiris (ECG applications)
Wilbert Burger (business management)
Redwan Ali (volunteer computing)

Peter den Brock (security)

Remko Kleinjan (initial pilot)
Redmar van Leeuwarden (initial pilot)
Sander van Vliet (initial pilot)

Cooperations

Nico Bruens, Erasmus MC, Rotterdam
Peter Walgemoed, Carelliance, Eindhoven
Prof. Dr. Ben Segal, CERN, Geneva
Prof. Dr. Ulrich Sax, MediGRID, Göttingen
Prof. Dr. Otto Riehnhoff, MediGRID, Göttingen
Dr. Maurice Bouhuis, SARA, Amsterdam

Steering Committee:

Prof. Dr. Frank G. Grosveld, ErasmusMC
Prof. Dr. Hans van den Berg, HR
Braam van Laar, ErasmusMC
Rob Ketelaar, HR

Institutions:

Erasmus Medical Center
Hogeschool Rotterdam
The German MediGRID and D-Grid Initiatives
The German Ministry for Science and Technology

Acknowledgem

Erasmus Computing Grid

Erasmus MC

HOGESCHOOL ROTTERDAM

Workforce:

Rob de Graaf (E

Anis Abu

Will

Co

Nico Bi

Peter Wa

Prof. Dr. Be

Prof. Dr. Ulri

Prof. Dr. Otto K

Dr. Maurice Boul

Future Ambition:

Griddify the Rotterdam Region &

The Netherlands

The Erasmus Computing Grid

–

Building a Super-Computer for FREE

Knoch, T. A. & de Zeeuw, L. V.

Erasmus Medical Center, Erasmus University of Rotterdam, Rotterdam, The Netherlands, May, 2007.

Abstract

Today advances in scientific research as well as clinical diagnostics and treatment are inevitably connected with information solutions concerning computation power and information storage. The needs for information technology are enormous and are in many cases the limiting factor for new scientific results or clinical diagnostics and treatment. At the Hogeschool Rotterdam and the Erasmus MC there is a massive need for computation power on a scale of 10,000 to 15,000 computers equivalent to ~20 to ~30 Tflops (10^{12} floating point operations per second) for a variety of work areas ranging from e.g. MRI and CT scan and microscopic image anlysis to DNA sequence analysis, protein and other structural simulations and analysis. Both institutions have already 13,000 computers, i.e. ~18 Tflops of computer power, available!

To make the needed computer power accessible, we started to build the Erasmus Computing Grid (ECG), which is connecting local computers in each institution via central management systems. The system guaranties security and any privacy rules through the used software as well as through our set-up and a NAN and ISO certification process being under way. Similar systems run already world-wide on entire institutions including secured environments like government institutions or banks. Currently, the ECG has a computational power of ~5 Tflops and is one of or already the largest desktop grid in the world. At the Hogeschool Rotterdam meanwhile all computers were included in the ECG. Currently, 10 departments with ~15 projects at the Erasmus MC depend on using the ECG and are preparing or prepared their analysis programs or are already in production state. The Erasmus Computing Grid office and an advisory and control board were set-up.

To sustain the ECG now further infrastructure measures have to be taken. Central hardware and specialist personal needs to be put in place for capacity, security and usability reasons for the application at Erasmus MC. This is also necessary in respect to NAN and ISO certification towards diagnostic and commercial ECG use, for which there is great need and potential. Beyond the link to the Dutch BigGrid Initiative and the German MediGRID should be prepared for and realized due to the great interest for cooperation. There is also big political interest from the government to relieve the pressure on computational needs in The Netherlands and to strengthen the Dutch position in the field of high performance computing. In both fields the ECG should be brought into a leading position by establishing the Erasmus MC a centre of excellence for high-performance computing in the medical field in respect to Europe and world-wide.

Consequently, we successfully started to build a super-computer at the Hogeschool Rotterdam and Erasmus MC with great opportunities for scientific research, clinical diagnostics and research as well as student training. This will put both institutions in the position to play a major world-wide role in high-performance computing. This will open entire new possibilities for both institutions in terms of recognition and new funding possibilities and is of major importance for The Netherlands and the EU.

Corresponding author email contact: TA.Knoch@taknoch.org

Keywords:

Genome, genomics, genome organization, genome architecture, structural sequencing, architectural sequencing, systems genomics, coevolution, holistic genetics, genome mechanics, genome function, genetics, gene regulation, replication, transcription, repair, homologous recombination, simultaneous co-transfection, cell division, mitosis, metaphase, interphase, cell nucleus, nuclear structure, nuclear organization, chromatin density distribution, nuclear morphology, chromosome territories, subchromosomal domains, chromatin loop aggregates, chromatin rosettes, chromatin loops, chromatin fibre, chromatin density, persistence length, spatial distance measurement, histones, H1.0, H2A, H2B, H3, H4, mH2A1.2, DNA sequence, complete sequenced genomes, molecular transport, obstructed diffusion, anomalous diffusion, percolation, long-range correlations, fractal analysis, scaling analysis, exact yard-stick dimension, box-counting dimension, lacunarity dimension, local nuclear dimension, nuclear diffuseness, parallel super computing, grid computing, volunteer computing, Brownian Dynamics, Monte Carlo, fluorescence in situ hybridization, confocal laser scanning microscopy, fluorescence correlation spectroscopy, super resolution microscopy, spatial precision distance microscopy, auto-fluorescent proteins, CFP, GFP, YFP, DsRed, fusion protein, in vivo labelling, information browser, visual data base access, holistic viewing system, integrative data management, extreme visualization, three-dimensional virtual environment, virtual paper tool.

Literature References

Knoch, T. A. Dreidimensionale Organisation von Chromosomen-Domänen in Simulation und Experiment. (Three-dimensional organization of chromosome domains in simulation and experiment.) *Diploma Thesis*, Faculty for Physics and Astronomy, Ruperto-Carola University, Heidelberg, Germany, 1998, and TAK Press, Tobias A. Knoch, Mannheim, Germany, ISBN 3-00-010685-5 and ISBN 978-3-00-010685-9 (soft cover, 2rd ed.), ISBN 3-00-035857-9 and ISBN 978-3-00-0358857-0 (hard cover, 2rd ed.), ISBN 3-00-035858-7, and ISBN 978-3-00-035858-6 (DVD, 2rd ed.), 1998.

Knoch, T. A., Münkel, C. & Langowski, J. Three-dimensional organization of chromosome territories and the human cell nucleus - about the structure of a self replicating nano fabrication site. *Foresight Institute - Article Archive*, Foresight Institute, Palo Alto, *CA, USA,* http://www. foresight.org, 1- 6, 1998.

Knoch, T. A., Münkel, C. & Langowski, J. Three-Dimensional Organization of Chromosome Territories and the Human Interphase Nucleus. *High Performance Scientific Supercomputing*, editor Wilfried Juling, Scientific Supercomputing Center (SSC) Karlsruhe, University of Karlsruhe (TH), 27- 29, 1999.

Knoch, T. A., Münkel, C. & Langowski, J. Three-dimensional organization of chromosome territories in the human interphase nucleus. *High Performance Computing in Science and Engineering 1999*, editors Krause, E. & Jäger, W., High-Performance Computing Center (HLRS) Stuttgart, University of Stuttgart, Springer Berlin-Heidelberg-New York, ISBN 3-540-66504-8, 229-238, 2000.

Bestvater, F., **Knoch, T. A.**, Langowski, J. & Spiess, E. GFP-Walking: Artificial construct conversions caused by simultaneous cotransfection. *BioTechniques* 32(4), 844-854, 2002.

Knoch, T. A. (editor), Backes, M., Baumgärtner, V., Eysel, G., Fehrenbach, H., Göker, M., Hampl, J., Hampl, U., Hartmann, D., Hitzelberger, H., Nambena, J., Rehberg, U., Schmidt, S., Weber, A., & Weidemann, T. Humanökologische Perspectiven Wechsel - Festschrift zu Ehren des 70. Geburtstags von Prof. Dr. Kurt Egger. Human Ecology Working Group, Ruperto-Carola University of Heidelberg, Heidelberg, Germany, 2002.

Knoch, T. A. Approaching the three-dimensional organization of the human genome: structural-, scaling- and dynamic properties in the simulation of interphase chromosomes and cell nuclei, long- range correlations in complete genomes, *in vivo* quantification of the chromatin distribution, construct conversions in simultaneous co-transfections. *Dissertation*, Ruperto-Carola University, Heidelberg, Germany, and TAK†Press, Tobias A. Knoch, Mannheim, Germany, ISBN 3-00-009959-X and ISBN 978-3-00-009959-5

(soft cover, 3rd ed.), ISBN 3-00-009960-3 and ISBN 978-3-00-009960-1 (hard cover, 3rd ed.), ISBN 3-00-035856-9 and ISBN 978-3-00-010685-9 (DVD, 3rd ed.) 2002.

Knoch, T. A. Towards a holistic understanding of the human genome by determination and integration of its sequential and three-dimensional organization. *High Performance Computing in Science and Engineering 2003*, editors Krause, E., Jäger, W. & Resch, M., High-Performance Computing Center (HLRS) Stuttgart, University of Stuttgart, Springer Berlin-Heidelberg-New York, ISBN 3- 540-40850-9, 421-440, 2003.

Wachsmuth, M., Weidemann, T., Müller, G., Urs W. Hoffmann-Rohrer, **Knoch, T. A.**, Waldeck, W. & Langowski, J. Analyzing intracellular binding and diffusion with continuous fluorescence photobleaching. *Biophys. J.* 84(5), 3353-3363, 2003.

Weidemann, T., Wachsmuth, M., **Knoch, T. A.**, Müller, G., Waldeck, W. & Langowski, J. Counting nucleosomes in living cells with a combination of fluorescence correlation spectroscopy and confocal imaging. *J. Mol. Biol.* 334(2), 229-240, 2003.

Fejes Tóth, K., **Knoch**, **T. A.**, Wachsmuth, M., Frank-Stöhr, M., Stöhr, M., Bacher, C. P., Müller, G. & Rippe, K. Trichostatin A induced histone acetylation causes decondensation of interphase chromatin. *J. Cell Science* 177, 4277-4287, 2004.

Ermler, S., Krunic, D., **Knoch, T. A.**, Moshir, S., Mai, S., Greulich-Bode, K. M. & Boukamp, P. Cell cycle-dependent 3D distribution of telomeres and telomere repeat-binding factor 2 (TRF2) in HaCaT and HaCaT-myc cells. *Europ. J. Cell Biol.* 83(11-12), 681-690, 2004.

Kost, C., Gama de Oliveira, E., **Knoch, T. A.** & Wirth, R. Spatio-temporal permanence and plasticity of foraging trails in young and mature leaf-cutting ant colonies (*Atta spp.*). *J. Trop. Ecol.* 21(6), 677- 688, 2005.

Winnefeld, M., Grewenig, A., Schnölzer, M., Spring, H., **Knoch, T. A.**, Gan, E. C., Rommelaere, J. & Cziepluch, C. Human SGT interacts with BAG-6/Bat-3/Scythe and cells with reduced levels of either protein display persistence of few misaligned chromosomes and mitotic arrest. *Exp. Cell Res.* 312, 2500-2514, 2006.

Sax, U., Weisbecker, A., Falkner, J., Viezens, F., Yassene, M., Hartung, M., Bart, J., Krefting, D., **Knoch, T. A.** & Semler, S. Grid-basierte Services für die elektronische Patientenakte der Zukunft. *E- HEALTH-COM - Magazin für Gesundheitstelematik und Telemedizin*, 4(2), 61-63, 2007.

de Zeeuw, L. V., **Knoch, T. A.,** van den Berg, J. & Grosveld, F. G. Erasmus Computing Grid - Het bouwen van een 20 TeraFLOP virtuelle supercomputer. *NIOC proceedings 2007 - het perspective of lange termijn*. editor Frederik, H. NIOC, Amsterdam, The Netherlands, 52-59, 2007.

Rauch, J., **Knoch, T. A.** , Solovei, I., Teller, K. Stein, S., Buiting, K., Horsthemke, B., Langowski, J., Cremer, T., Hausmann, M. & Cremer, C. Lightoptical precision measurements of the Prader- Willi/Angelman Syndrome imprinting locus in human cell nuclei indicate maximum condensation changes in the few hundred nanometer range. *Differentiation* 76(1), 66-82, 2008.

Sax, U., Weisbecker, A., Falkner, J., Viezens, F., Mohammed, Y., Hartung, M., Bart, J., Krefting, D., **Knoch, T. A.** & Semler, S. C. Auf dem Weg zur individualisierten Medizin - Grid-basierte Services für die EPA der Zukunft. *Telemedizinführer Deutschland 2008*, editor Jäckel, A. Deutsches Medizinforum, Minerva KG, Darmstadt, ISBN 3-937948-06-6, ISBN-13 9783937948065, 47-51, 2008.

Drägestein, K. A., van Capellen, W. A., van Haren, J. Tsibidis, G. D., Akhmanova, A., **Knoch, T. A.**, Grosveld, F. G. & Galjart, N. Dynamic behavior of GFP-CLIP-170 reveals fast protein turnover on microtubule plus ends. *J. Cell Biol. 180(4)*, 729-737, 2008.

Jhunjhunwala, S., van Zelm, M. C., Peak, M. M., Cutchin, S., Riblet, R., van Dongen, J. J. M., Grosveld, F. G., **Knoch, T. A.**$^{+}$ & Murre, C.$^{+}$ The 3D-structure of the Immunoglobulin Heavy Chain Locus: implications for long-range genomic interactions. *Cell 133(2)*, 265-279, 2008.

Krefting, D., Bart, J., Beronov, K., Dzhimova, O., Falkner, J., Hartung, M., Hoheisel, A., **Knoch, T. A.**, Lingner, T., Mohammed, Y., Peter, K., Rahm, E., Sax, U., Sommerfeld, D., Steinke, T., Tolxdorff, T., Vossberg, M.,

Viezens, F. & Weisbecker, A. MediGRID - Towards a user friendly secured grid infrastructure. *Future Generation Computer Systems 25(3)*, 326-336, 2008.