ScalaLife

Scalable Software Services
for Life Science

Improving the Usage of eInfrastructures for Life Science Researchers

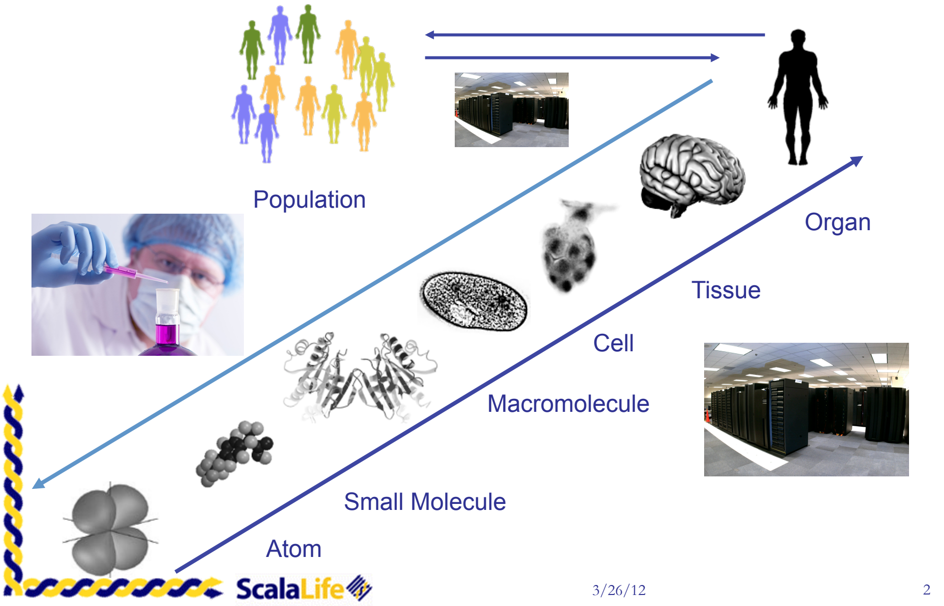
Erwin Laure
KTH

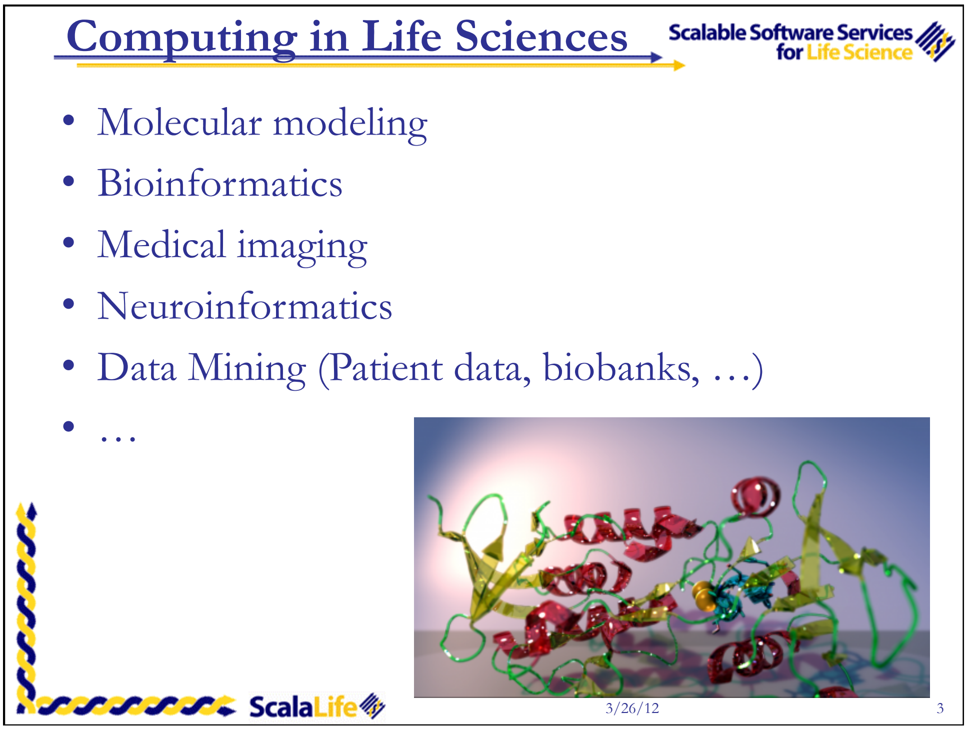
CAPACITIES

e-infrastructure

Life Science and Health

Scalable Software Services for Life Science

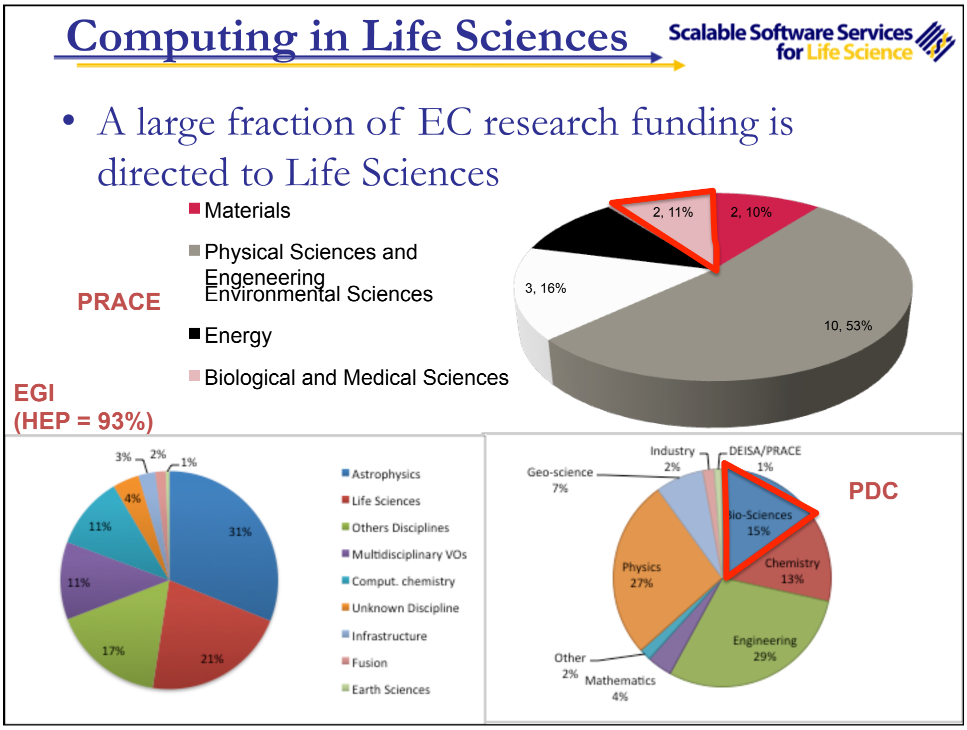




Computing in Life Sciences

- Molecular modeling
- Bioinformatics
- Medical imaging
- Neuroinformatics
- Data Mining (Patient data, biobanks, …)
- …

3/26/12 3



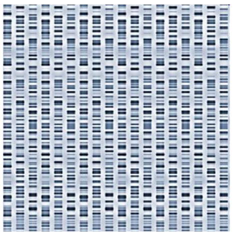
Computing in Life Sciences

- A large fraction of EC research funding is directed to Life Sciences

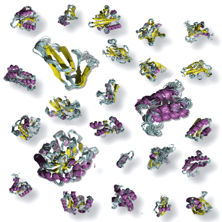
YET

Life Science is not a major user of eInfrastructures

Increasing Life Science Computations



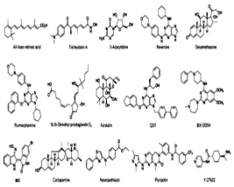
Genomes



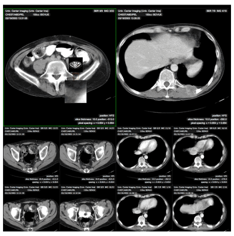
Structures

Integration!

Data



Molecules



Images

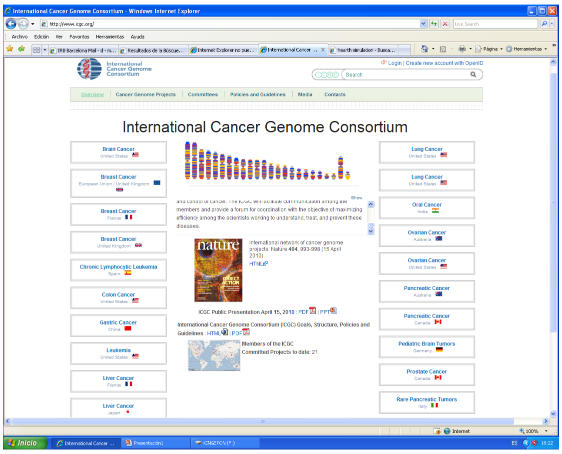
Slides Courtesy Modesto Orozco, IRB, EESI

International Cancer Genome Consortium

International Cancer Genome Project

Scalable Software Services for Life Science

Experimental data 1 patient < 1 day
Primary analysis 1 patient → 5 days!
Secondary analysis 1 patient → months!



International Cancer Genome Consortium

Brain Cancer United States	Lung Cancer United States
Breast Cancer European Union / United Kingdom	Lung Cancer United States
Breast Cancer France	Oral Cancer India
Breast Cancer United Kingdom	Ovarian Cancer Australia
Chronic Lymphocytic Leukemia Spain	Ovarian Cancer United States
Colon Cancer United States	Pancreatic Cancer Australia
Gastric Cancer China	Pancreatic Cancer Canada
Leukemia United States	Pediatric Brain Tumors Germany
Liver Cancer France	Prostate Cancer Canada
Liver Cancer Japan	Rare Pancreatic Tumors Italy

and control of cancer. The ICGC will facilitate communication among the members and provide a forum for coordination with the objective of maximizing efficiency among the scientists working to understand, treat, and prevent these diseases.

International network of cancer genome projects. Nature **464**, 993-998 (15 April 2010)
HTML

ICGC Public Presentation April 15, 2010 : PDF | PPT

International Cancer Genome Consortium (ICGC) Goals, Structure, Policies and Guidelines : HTML | PDF

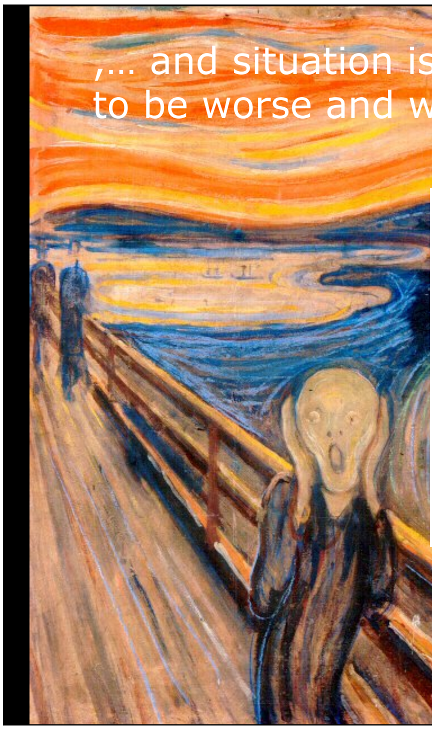
Members of the ICGC
Committed Projects to date: 21

50 cancers

25000 cancer genomes

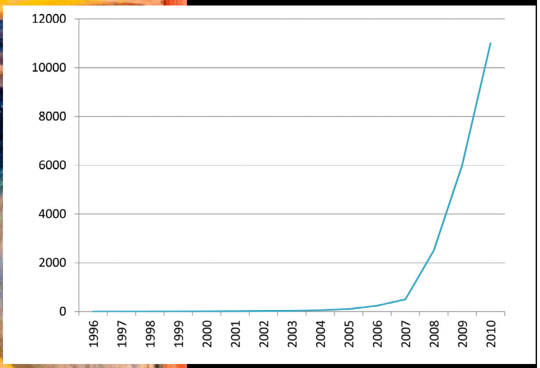
Puente et al., Nature 2011

ScalaLife

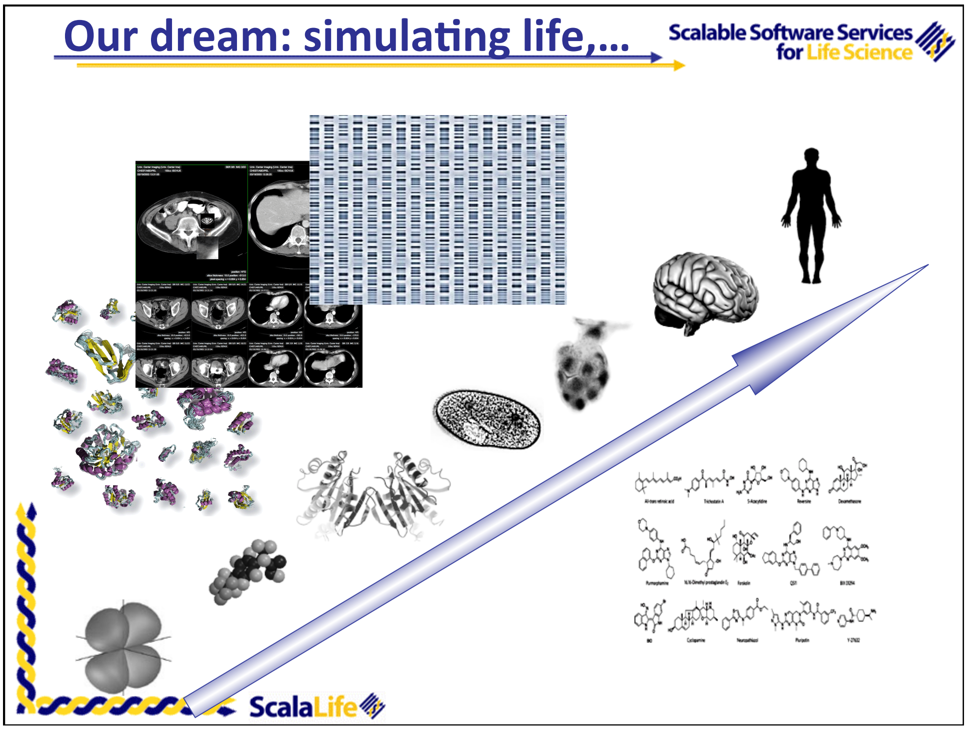


,... and situation is going to be worse and worse,...

ELIXIR



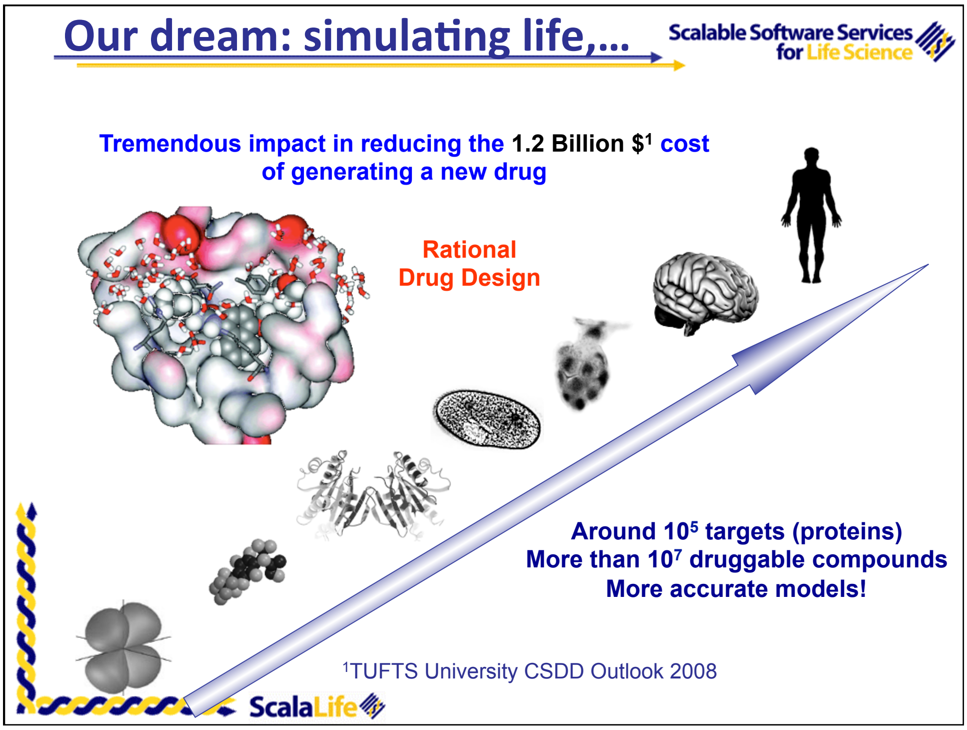
Data storage at EBI (Tb)
EBI 2010 annual scientific report



Our dream: simulating life,…

Scalable Software Services for Life Science

ScalaLife



Our dream: simulating life,…

Scalable Software Services for Life Science

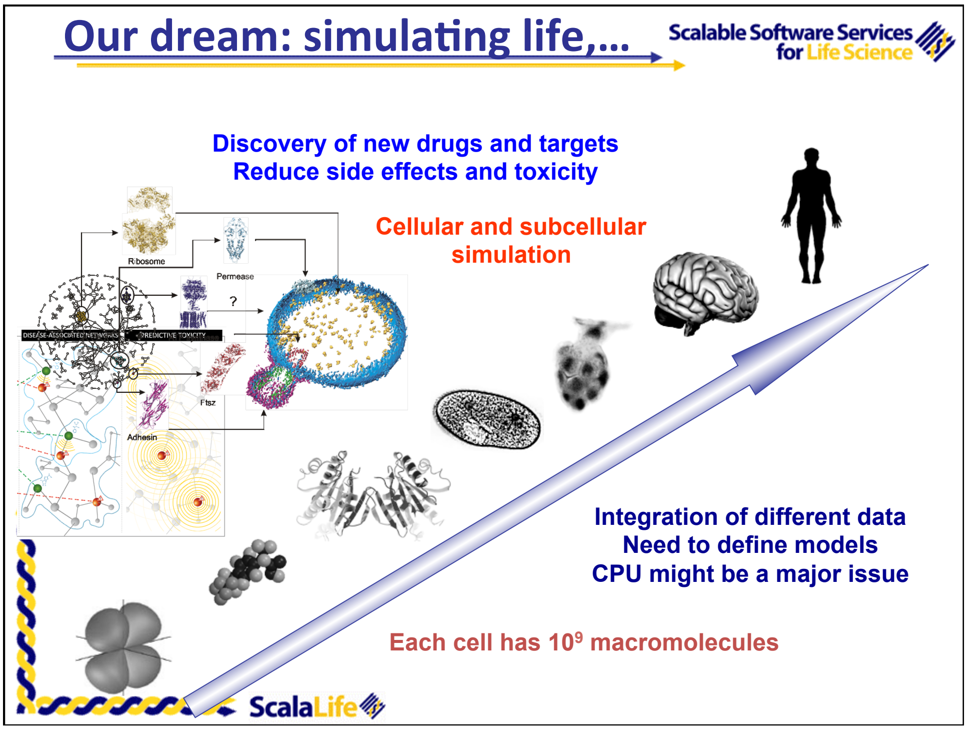
Tremendous impact in reducing the **1.2 Billion $**[1] cost of generating a new drug

Rational Drug Design

Around 10^5 targets (proteins)
More than 10^7 druggable compounds
More accurate models!

[1]TUFTS University CSDD Outlook 2008

ScalaLife



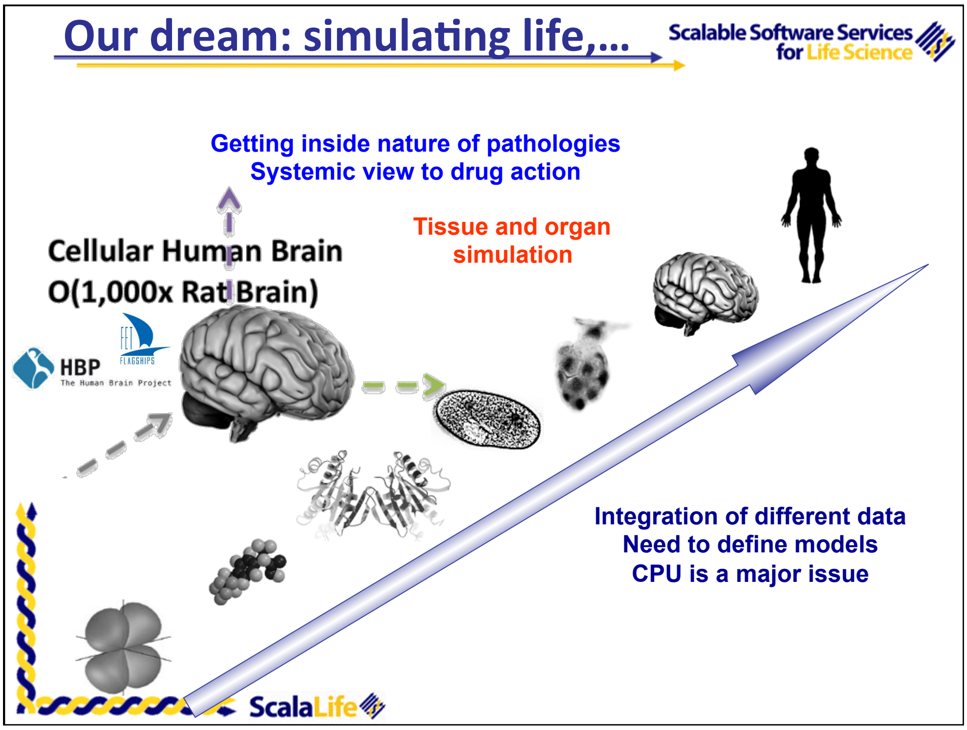
Our dream: simulating life,…

Discovery of new drugs and targets
Reduce side effects and toxicity

Cellular and subcellular
simulation

Integration of different data
Need to define models
CPU might be a major issue

Each cell has 10^9 macromolecules



Our dream: simulating life,…

Getting inside nature of pathologies
Systemic view to drug action

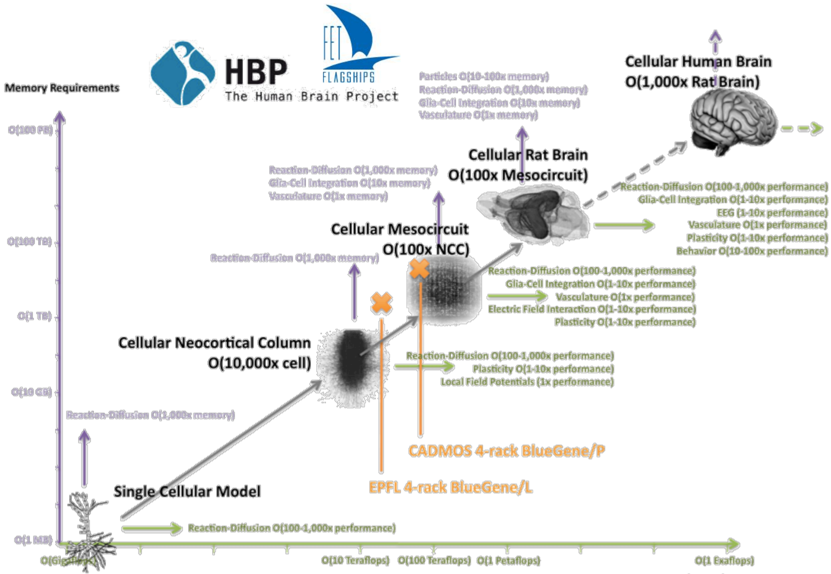
Tissue and organ
simulation

Cellular Human Brain
O(1,000x Rat Brain)

Integration of different data
Need to define models
CPU is a major issue

FET: Human Brain Project

Scalable Software Services for Life Science



F Schürmann, H Markram (Blue Brain Project, EPFL)

ScalaLife

Major European Efforts

Scalable Software Services for Life Science

- ELIXIR, Bioinformatics community
- EMBL-EBI
- INSTRUCT
- Research consortia: ENCODE, ICGC, iHEC
- FET Flagships: Personalized Medicine, Human Brain Project.
- EGI, PRACE
- Pharmaceutical Industry
- Computational Biology and Bioinformatics Software

Life-Science needs Exascale

ScalaLife

Main conclusions of the EESI Life Science panel

- Life science requirements for HPC are enormous and grow exponentially. There is not question that Life Sciences community has ExaScale needs.
- Lack of HPC resources will simply kill several of the most important projects in Life Sciences, including several EU Flagship ones (Human Brain, Personalized Medicine).
- Many other issues rather than FLOPS are important in Life Sciences, often more important.
- **Exascale** should be the keyword, not ExaFlop.
- The experts propose Co-development programs to deploy Exascale to Life Sciences

---

Scalable Software Services for Life Science

But what's the typical situation of a Researcher in Life Science today?

ScalaLife

What Software to Use?

AMBER

GROMACS

HOOMD

Tinker

NAMD

LAMMPS

ESPResSo

COSMOS

Presto

CHARMM

MacroModel

MDynaMix

ACEMD

Kalypso

ADUN

Materials Studio

Desmond

Ascalaph

GULP

LPMD

Hippo

RedMD

MOLDY

Culgi

MOSCITO

ProtoMol

GROMOS

YASARA

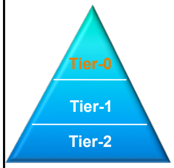
Abalone

XMD

ORAC

Where to run?

PRACE Systems?



First production system available: **1 Petaflop/s** IBM BlueGene/P (**JUGENE**) at GCS (Gauss Centre for Supercomputing) partner FZJ (Forschungszentrum Jülich)

Italy and Spain to deploy Tier-0 systems in mid 2012.

Second production system available: Bull Bullx **CURIE** at GENCI partner CEA. Full capacity of **1.8 Petaflop/s** reached by late 2011.

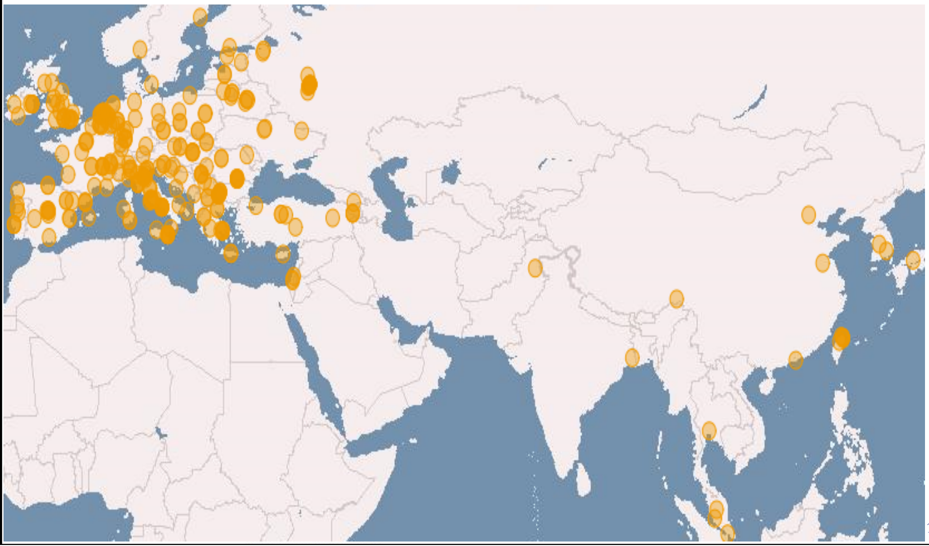
Third production system available by the end of 2011: **1 Petaflop/s** Cray (**HERMIT**) at GCS partner HLRS (High Performance Computing Center Stuttgart). Upgrade to 4-5 Petaflop/s planned in 2013.

Fourth production system available by mid 2012: **3 Petaflop/s** IBM (**SuperMUC**) at GCS partner LRZ (Leibniz-Rechenzentrum).

Where to run?

Scalable Software Services for Life Science

EGI?



Where to run?

Scalable Software Services for Life Science

Your local computing center?

Where to run?

On the Cloud?

And More Questions …

Which architecture?

I need extra functionality in Y software, who can implement?

How to get help with PRACE?

Where to run?

How to get access to the X system?

How can I run on EGI?

Which software fits best to my problem?

What is the best way of compiling and running the specific Y code on X system?

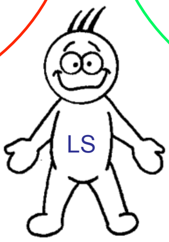
I want to use GPGPUs for software Y, but how?

My application is not scaling, what to do?

The expert at center X needs help with software Y , can you help him/her?

What is the best way of data handling for my problem?

What is performance analysis and who/how to do it?

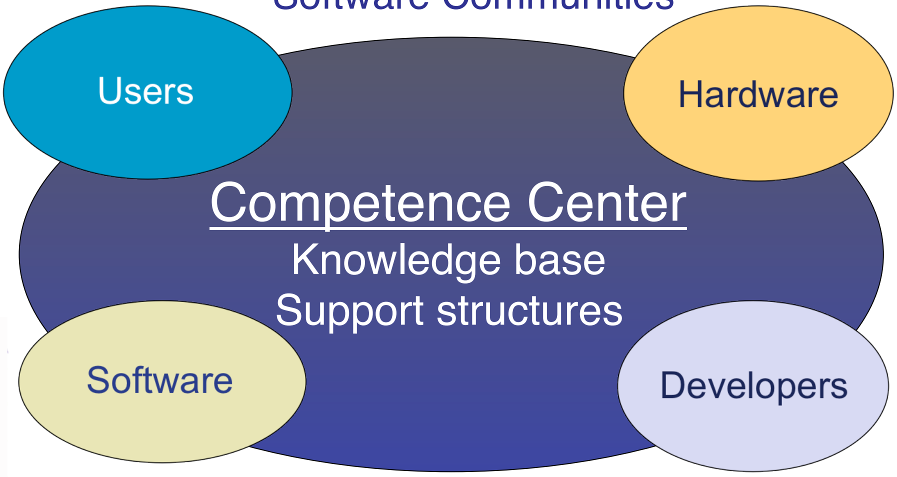


ScalaLife provides

- A **Competence Centre** as one-stop-shop for the computational Life Science community providing
  - **Optimized simulation codes**
  - **Advise on usage of e-Infrastructure**
  - **Expert support**
- The Competence Centre extends much beyond the lifetime of ScalaLife and is expected to turn into a **long lasting structure**

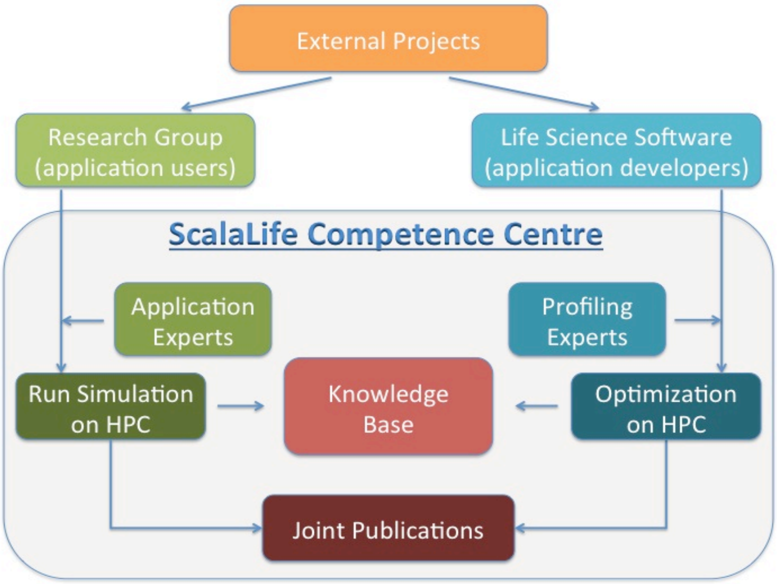
Competence Center

“**One-stop-shop**” for Life Science Software Communities



www.ScalaLife.eu

Competence Centre Structure



External Projects

Research Group (application users)

Life Science Software (application developers)

ScalaLife Competence Centre

Application Experts

Profiling Experts

Run Simulation on HPC

Knowledge Base

Optimization on HPC

Joint Publications

ScalaLife delivers

- Hierarchical **parallelization** of key European open source codes
- **Ensemble & high-throughput** techniques for multi-core and streaming architectures to parallelize in the problem domain
- Establish **open software standards** for data storage and exchange
- Implement, document, and maintain such techniques in pilot **European open-source codes** – using licenses that allow code reuse

ScalaLife and EGI

- ScalaLife delivers training to EGI via training market place; partly in collaboration with we-nmr
- Help making software available
- Collaborate with LSGC/HealthGrid on supporting users
  - Exchange of expertise, information, and user support

ScalaLife and PRACE

- Provide support PRACE/DECI Life Science users
- Provide software and support to PRACE centers
- Provide benchmark results to help with evaluation of applications
- Collaborate on (peta)scaling of Dalton and GROMACS

ScalaLife

Scalable Software Services for Life Science

Case Studies

e-infrastructure

CAPACITIES

GPUs and FPGAs

Scalable Software Services for Life Science

- What impact can new hardware trends have?
- GPUs have a great promise – but can they really be used?
- What about FPGAs – should one invest time into them?

ScalaLife

Feasibility study, Gromacs on FPGA

Scalable Software Services for Life Science

- The non-bonded forces of GROMACS has been ported to FPGAs.
- This non-bounded forces calculation covers roughly 70% of the execution time.
  - Maximum speed-up 3X.

Gene Amdahl

- Assumes each node has a CPU and one or many FPGAs connected over PCI-E .

Gromacs Results

Scalable Software Services for Life Science

Rough initial results (Mega atom pair interactions/s)

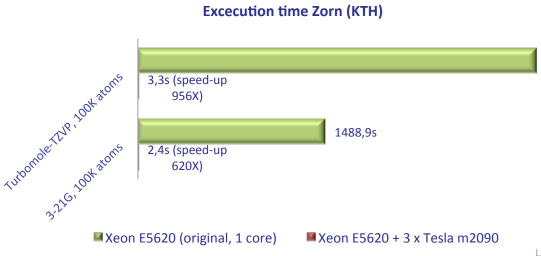
- 300 M/s - 4 core CPU with hyperthreading
- 960 M/s - 4 core CPU with hyperthreading, using SSE
- 3125 M/s - using an C2075 GPU
- 4000 M/s - AC3611 (medium size Altera FPGA)
- Theoretically 32 000 M/s - Convey system, HC-1ex (4 very large Xilinx FPGAs, room for 2 modules per FPGA).

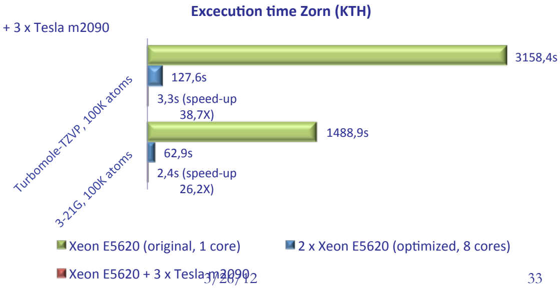
Performance:

- We need approximately 2 Gbyte/s for each module, the PCI-E connection would be able to handle 2 modules.
- The HC-1ex is not limited by the PCI-E connection.
- Need to accelerate outside the non-bonded forces to get a speed-up better than 3X. (use one FPGA to non bonded forces and one FPGA to other calculations)

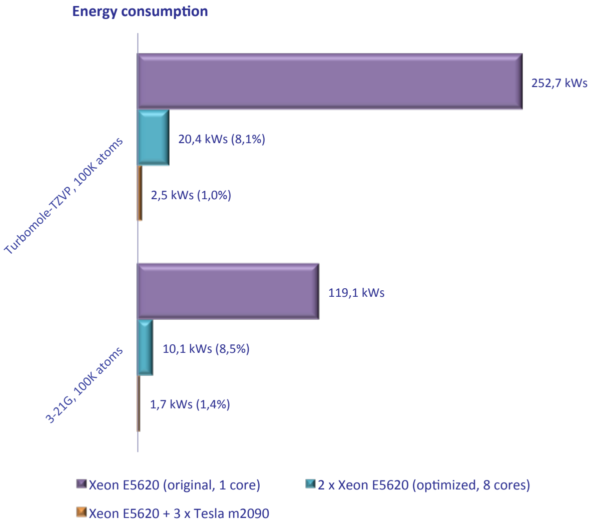
GPUs

- Atomic simulation using ERGO
- Using two different basis sets





Energy Consumption

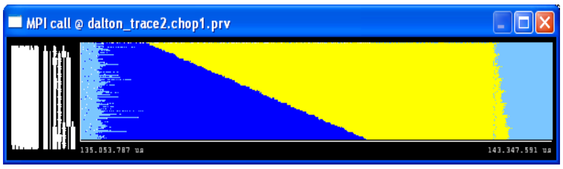


Findings

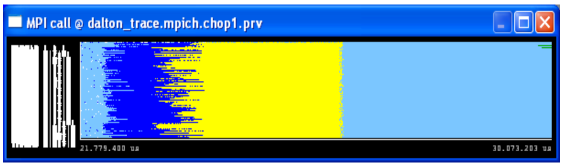
- FPGAs seems to be a suitable platform for molecular dynamics simulations
  - Expertise in FPGAs is limited
    - Hard to find experienced engineers
- GPUs provide some additional speedup over optimized multicore
  - But at expense of portability and coding effort
- Load balancing is a big issue in hybrid architectures (CPU and accelerator)

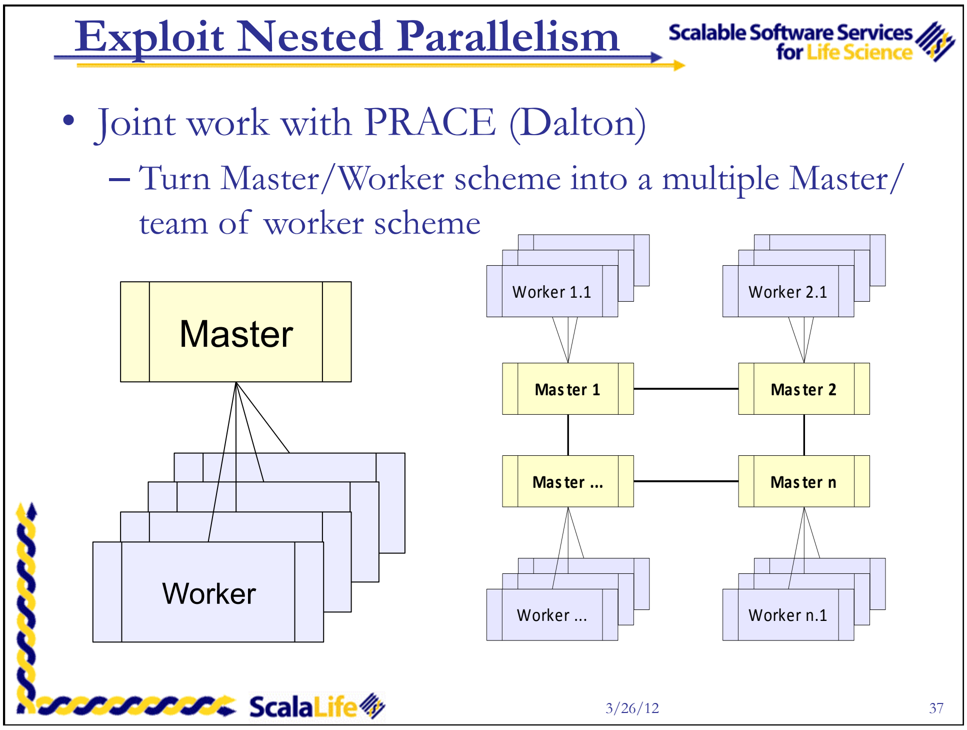
Performance Tools

- Use Performance Tools to identify bottlenecks and improve program structure
- E.g. Master/Slave with MPI_Recv_Any:
  - OPENMPI



  - MVAPICH2



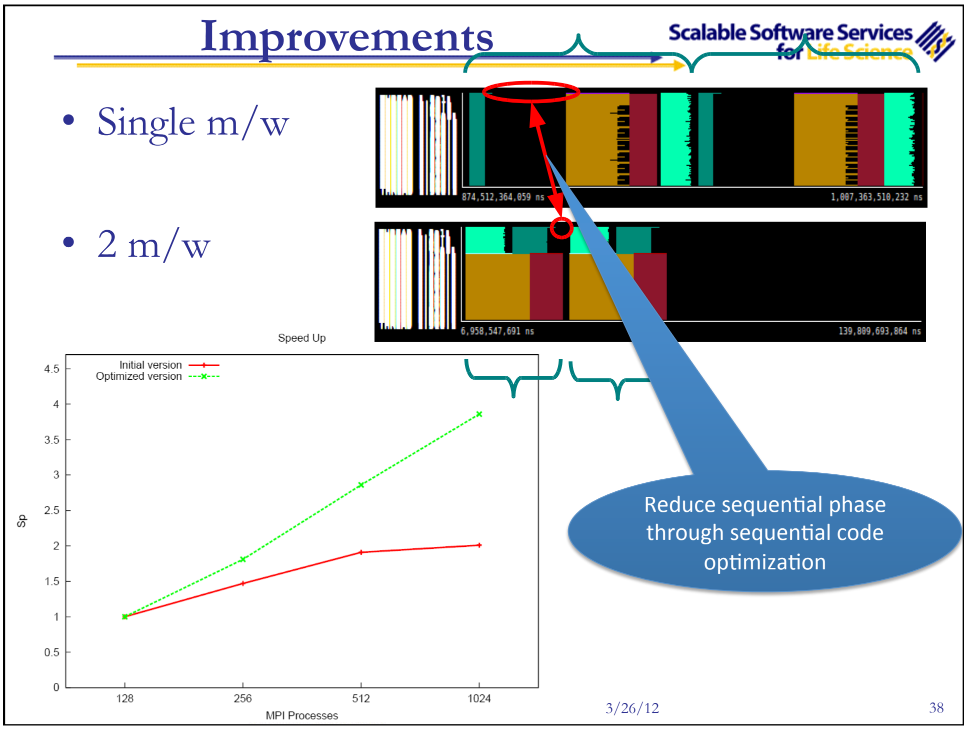


Exploit Nested Parallelism

Scalable Software Services for Life Science

- Joint work with PRACE (Dalton)
  - Turn Master/Worker scheme into a multiple Master/ team of worker scheme

ScalaLife



Improvements

Scalable Software Services for Life Science

- Single m/w
- 2 m/w

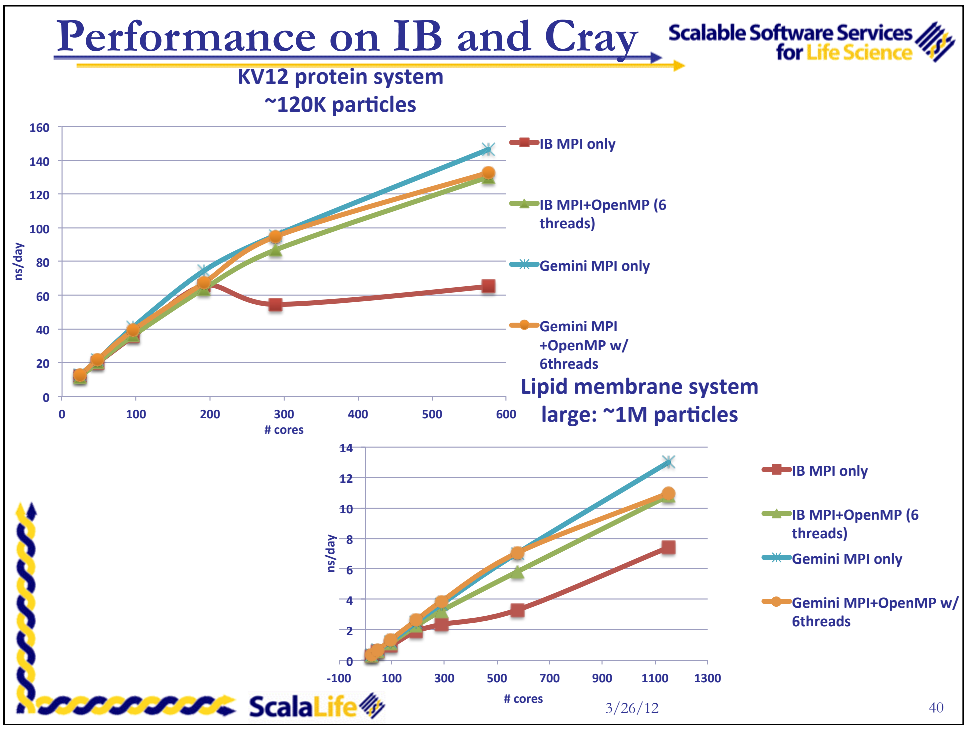
Reduce sequential phase through sequential code optimization

Hybrid Parallelization

Scalable Software Services for Life Science

- Explore the effect of combined MPI and OpenMP parallelization on multi-core architectures
- Collaboration with PRACE
- Parallelization of GROMACS particle mesh Ewald (PME) method

ScalaLife 3/26/12 39



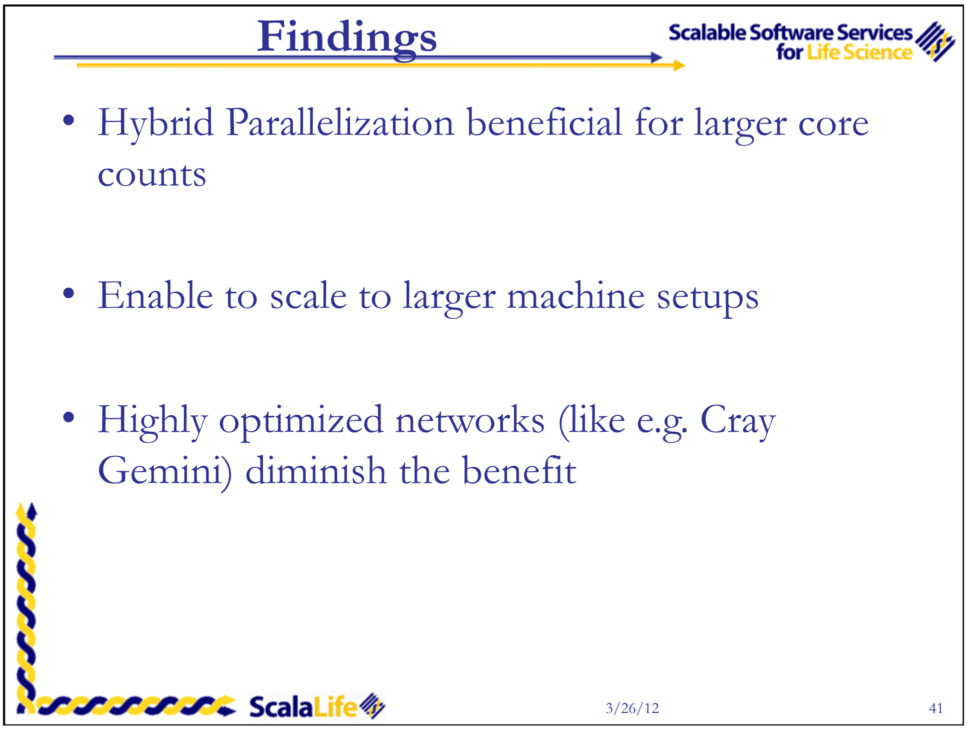
Performance on IB and Cray

Scalable Software Services for Life Science

KV12 protein system
~120K particles

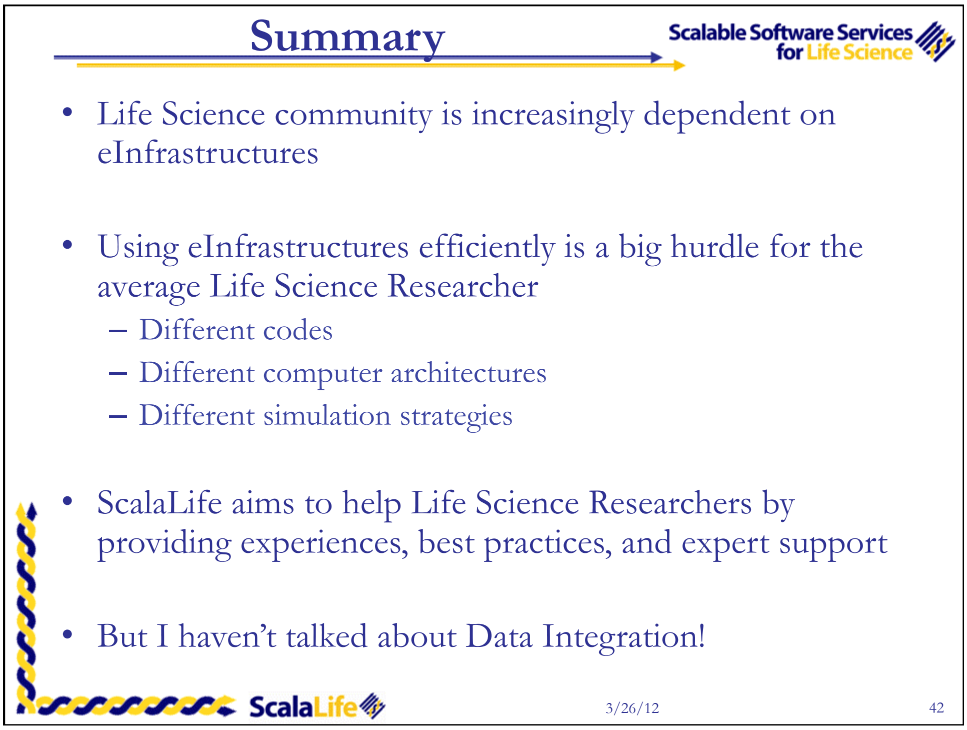
Lipid membrane system
large: ~1M particles

ScalaLife 3/26/12 40



Findings

- Hybrid Parallelization beneficial for larger core counts
- Enable to scale to larger machine setups
- Highly optimized networks (like e.g. Cray Gemini) diminish the benefit



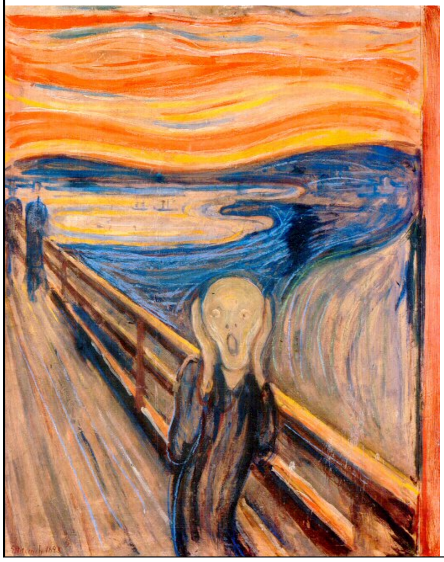
Summary

- Life Science community is increasingly dependent on eInfrastructures
- Using eInfrastructures efficiently is a big hurdle for the average Life Science Researcher
  - Different codes
  - Different computer architectures
  - Different simulation strategies
- ScalaLife aims to help Life Science Researchers by providing experiences, best practices, and expert support
- But I haven’t talked about Data Integration!

Help LifeScience Researchers to go

Scalable Software Services for Life Science

from today's situation



to smiling eInfrastructure users