IT-infrastruktura pro biologický a biomedicínský výzkum

Proč budovat v biologickém výzkumném ústavu (ÚŽFG) robustní IT infrastrukturu

- Současné biologické experimentální přístupy představují ***data intensive research***
- *Příklady:*
  - *Jeden standardní experiment s využitím **mikroskopie živých buněk** čítající 2-4 experimentální skupiny se pohybuje ve velikosti 20-100 GB*
  - ***Proteomové analýzy** na hmotnostním spektromeru produkují cca 2 000 GB (2TB) za týden*
  - **Next Generation Sequencing**: IlluminaHiSeq2000 produkuje cca 20 GB/den

- Tyto data je třeba nějakým způsobem ukládat (a také zálohovat), ale hlavně je efektivně analyzovat

Jak IT-infrastruktura pro takovéto požadavky vypadá

Cesnet - Internet

výkonné pracovní stanice

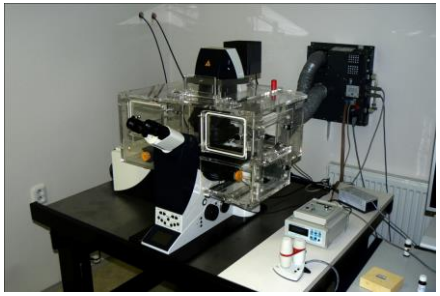
servery

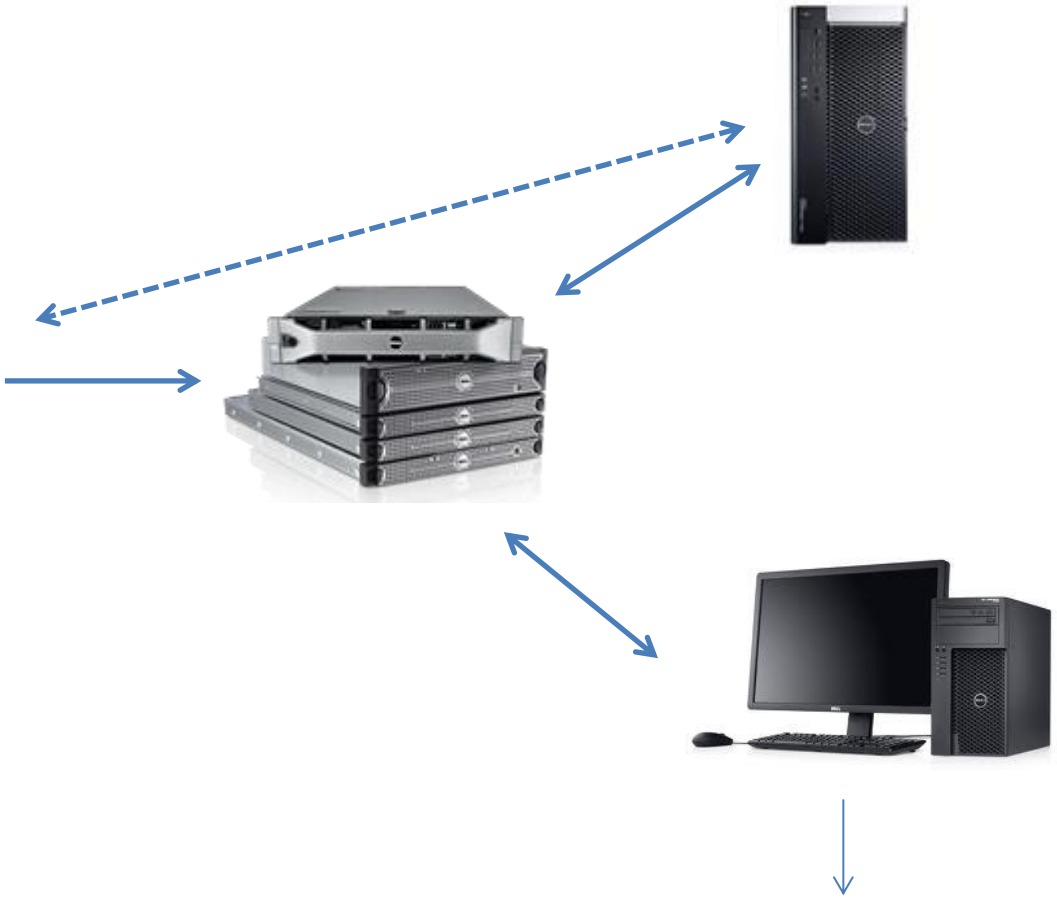
Síťové datové úložiště

stolní pracovní stanice

- Základní charakteristikou je oddělení ukládání a zpracování dat
- Většina analýz neprobíhá na stolních stanicích, ty se používají pouze pro ovládání
- Základem celku je **fungující rychlá síť, která však nespojuje pouze pracovní stanice, ale dělá z jednotlivých komponent jeden funkční celek**

Ukázka toku a zpracování dat – live cell imaging





výsledek, publikace

E-infrastruktura

Pojmem e-infrastruktura bývá označována komplexní sada informatických nástrojů použitelných pro řešení problémů z celé řady oborů. Použitelnost jejích služeb se neomezuje na přírodní vědy, jako je matematika, fyzika, chemie či informatika, ale zahrnuje i vědy humanitní a umění. S pokračující digitalizací dalších a dalších materiálů a rostoucím významem komunikačních technologií lze v současnosti jen stěží najít obor, kterému by neměla co nabídnout.

E-infrastruktura CESNET

Klíčovou složkou aktivit sdružení CESNET je rozvoj stejnojmenné e-infrastruktury. Jedná se o komplexní národní IT infrastrukturu určenou pro potřeby české vědy, výzkumu, vývoje a vzdělávání. Je zařazena do *Cestovní mapy ČR velkých infrastruktur pro výzkum, experimentální vývoj a inovace* Ministerstva školství, mládeže a tělovýchovy.

Mezi uživatelskými institucemi jsou tuzemské vysoké školy, ústavy Akademie věd České republiky, řada vědeckých a výzkumných institucí, ale i nemocnice, knihovny, střední školy a mnohé další.

Komponenty e-infrastruktury CESNET

Vzhledem k rozsahu naší e-infrastruktury je výhodné vymezit její logicky ucelené součásti a popisovat je odděleně. Základními komponentami jsou:

- komunikační infrastruktura (síť CESNET2),
- gridová infrastruktura pro náročné výpočty,
- infrastruktura datových úložišť a
- infrastruktura pro vzdálenou spolupráci

Klíčové projekty

Pro rozvoj e-infrastruktury CESNET jsou rozhodující zejména dva velké projekty, které poskytují nejvýznamnější zdroj financování:

- Projekt Velká infrastruktura CESNET
- Rozšíření národní informační infrastruktury pro výzkum a vývoj v regionech“ (projekt eIGeR)

Poslední změna: 15.5.2013

Co se zatím podařilo realizovat?

- Máme nově rychlý bezdrátový link do sítě CESNET (nyní stabilních 80 Mb/s, dříve velmi kolísavých 20 Mb/s)
- Během léta 2014 budeme mít připojení do CESNETU pomocí optiky (1-10 Gb/s)
- Pro analýzu obrazu již nyní používáme výkonnou pracovní stanici Dell T7500
- V rámci centra Pigmod se během 2 týdnů budou instalovat **výkonné pracovní stanice** a **datové úložiště** se zálohovanou kapacitou 100 TB (ÚŽFG má nyní „úložiště“ s kapacitou 1TB)

Co se je třeba realizovat?

- Zásadním způsobem upgradovat současnou počítačovou síť ústavu
- Zajistit personálně takovou správu IT infrastruktury jako celku, která bude nově vybudované kapacity schopna efektivně spravovat, a která bude schopna pomoci s využitím velké E-infrastruktury Cesnetu. Je nereálné, že by si toto zajišťovali vědečtí pracovníci.

CERN for molecular biology?

Ewan Birney, Joint Associate Director of EMBL-EBI, examines the similarities and differences between EMBL and CERN

This September I visited CERN again, this time with a technical delegation from the EBI to meet with their 'big physics data' counterparts. Our generous hosts, Ian Bird, Bob Jones and several experimental scientists showed us a great day, and gave us an extended opportunity to understand their data flow in detail.

CERN is a marvellous place, and the experiments conducted there share some similarities with large-scale biology projects: the large-scale data flow, the many stages of analysis, and the need for solid metadata. But the differences between high-energy physics (HEP) at CERN and molecular biology at EMBL are considerable. For example, the LHC data is about one order of magnitude larger than molecular biology data – though our data doubling time (~1 year) is shorter than their basic data doubling time (~2 years).

> "Each data-intensive science will need to adopt, adapt and sometimes create its own custom solutions."
>
> – *Ewan Birney*

Another difference is that HEP data flow is more 'starburst' in shape, emanating from a few central sites to progressively broader groups. Molecular biology data has a more 'uneven bow-tie' topology, with thousands of data-producing sites feeding a small number of global archive sites, which distribute to 100 000s of researchers worldwide.

Although HEP is not uniform – the results for each experiment are different – there is a far more limited repertoire of types-of-things one might want to catch. In molecular biology, the incredible heterogeneity of life is simply awe-inspiring. So in addition to data-volume tasks in molecular biology, we also have fundamental, large-scale data-structuring tasks.

What we can learn from CERN

There is a lot more we can learn in biology from HEP than one might expect. Some relates to pragmatic information engineering and some to deeper scientific aspects. There is certainly much to learn from how the LHC handles its data storage, but we should also look carefully at how they have created portable computer schemes.

There is a lot of knowledge we can share as well, for example in ontology engineering. The Experimental Factor Ontology's ability to deal with hundreds of component ontologies without exploding, could well be translated to other areas of science, and I think they were quietly impressed with the way we are still able to make good use of archived experimental data from the 1970s and 1980s. In molecular biology, I think this on-going use of data is something to be proud of.

Engaging further with our counterparts in HEP is something I am really looking forward to. It will be great to see Ian, Bob and the team at EMBL-EBI next year. CERN is a leader in data-intensive science, but each data-intensive science will need to adopt, adapt and sometimes create its own custom solutions.

✏ For a longer version of this article, visit Ewan's blog, 'Bioinformatician at large': bit.ly/1akFbRR

The Compact Muon Solelnoid at the Large Hadron Collider. CERN produces enough data in a year to fill CDs stacked several times the height of Mount Everest

In the presence of greatness

It began with great expectations, and just a sprinkling of nerves. Inspired by the Lindau Nobel Laureate Meeting, the first ever Heidelberg Laureate Forum had set the bar high, aiming to bring together young researchers in maths and computer science with some of the superstars of the fields – and, on 27 September, EMBL Heidelberg hosted the closing day of what was widely hailed as a fascinating week.

Amidst the spectacle of clicking cameras, lively discussions, and rows of personal chauffeurs waiting outside, talks placed strong emphasis on the role that interdisciplinary research could play in the future of computing, and many laureates detailed work that applied mathematical and computational approaches to biology.

The forum, which took place at institutions across the city, brought together 40 recipients of the Abel Prize and Fields Medal (mathematics) and the Turing Award (computer science), with more than 200 talented young researchers from across the globe. The aim was to emulate for maths and computer science what the Lindau meeting achieves for physics, chemistry and the life sciences, in providing a platform for dialogue between different scientific generations. A number of participants also took the chance to tour the lab and find out more about EMBL researchers working in computer science.

"Many young researchers I spoke with are shifting to look at problems that lie at traditional disciplinary boundaries," explains Amanda Randels, a fellow at Lawrence Livermore National Library. "The beginning and ending of the forum's scientific programme were absolutely perfect – from Raj Reddy's talk about who invented computing, to John Hopecroft's reflections on the future," adds Dana Mackenzie, a US-based science writer.

Turing Award Winner in 2003, Alan Kay, shares insights with the next generation of number crunchers

Rupert Lück, Michael Wahlers, Andrés Lindau, Tobias Rausch, Vladimír Beneš, Jonathon Blake and Markus Fritz who worked on the project

High performance computing for your NGS data

Since earlier this year, the expanded EMBL High Performance Computing (HPC) cluster – a group of linked nodes processing biological data in parallel – provides access to more than 3500 CPU cores and 32 TB of total memory. Supported by LSF, a new job scheduling system, it enables more efficient processing of the many computing requests the infrastructure receives every day. The infrastructure, developed by IT Services, benefits more than 100 users of the HPC facility. Recently, a big effort was made to improve the overall speed in processing the large amounts of next generation sequencing data produced by EMBL's Genomic Core Facility (GeneCore).

It is a growing problem in sequencing facilities around the world: how do you keep pace with the computing burden arising from the huge expansion in genetic sequencing? Next generation sequencing (NGS) does not produce a complete genome that researchers can read like a book; rather it generates something more reminiscent of a pile of shredded documents without any organisation of the fragments. At GeneCore one solution has come from a combination of more powerful computers and smarter use of these resources. "The continuous increase of our facility's NGS platform data output meant we effectively needed to put more machines and much more memory for our high throughput bioinformatics pipelines on the job," says Jonathon Blake, a bioinformatician in the facility. "Migrating workloads to the upgraded cluster has allowed us to clear bottlenecks producing a threefold increase in our processing throughput," adds Vladimír Beneš, head of GeneCore.

"It has produced a threefold increase in our processing throughput"

– Vladimír Beneš

"By working together with GeneCore staff, including bioinformaticians Tobias Rausch and Markus Fritz, as well as computational experts from many other groups, we have delivered an infrastructure that works: ultimately scientists are able to build upon a much more reliable HPC infrastructure to deliver fast, reliable results in a way that is highly conducive to successful science – at least until the next generation of sequencing machines comes along!" says Rupert Lück, head of IT Services.

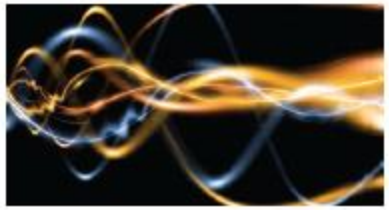
Research highlights

What makes you, you?

A study, conducted by scientists from nine institutes, including EMBL-EBI, presented a map that points to the genetic causes of differences between people. Published in *Nature* and *Nature Biotechnology*, the work offers the largest ever dataset linking human genomes to gene activity at the level of RNA.

Through the noise

Scientists in the Huber group at EMBL Heidelberg have pinpointed key activities in the human genome that are important for the understanding of health and disease. The findings, published in *PNAS* in September, highlight finely tuned, crucial events within a seemingly chaotic landscape.

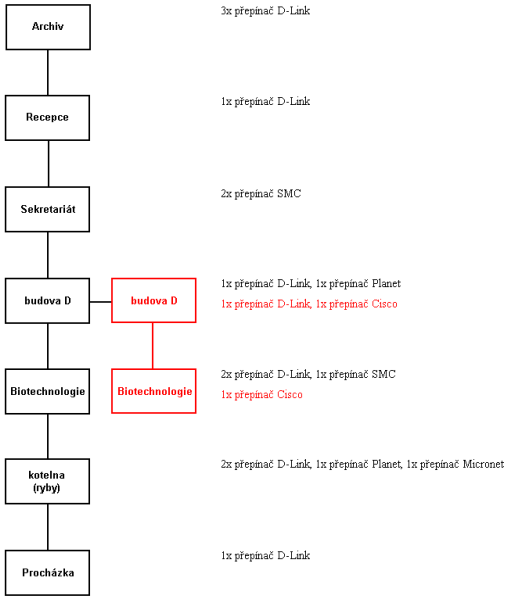


Děkuji za pozornost.

Schéma rozvaděčů počítačové sítě AVČR Liběchov

stav 29. srpna 2013

rychlost sítě: 100 Mb/s
rychlost sítě: 1000 Mb/s




Sun
IBM
PROMISE VTrak M610i