BioIT World 2015

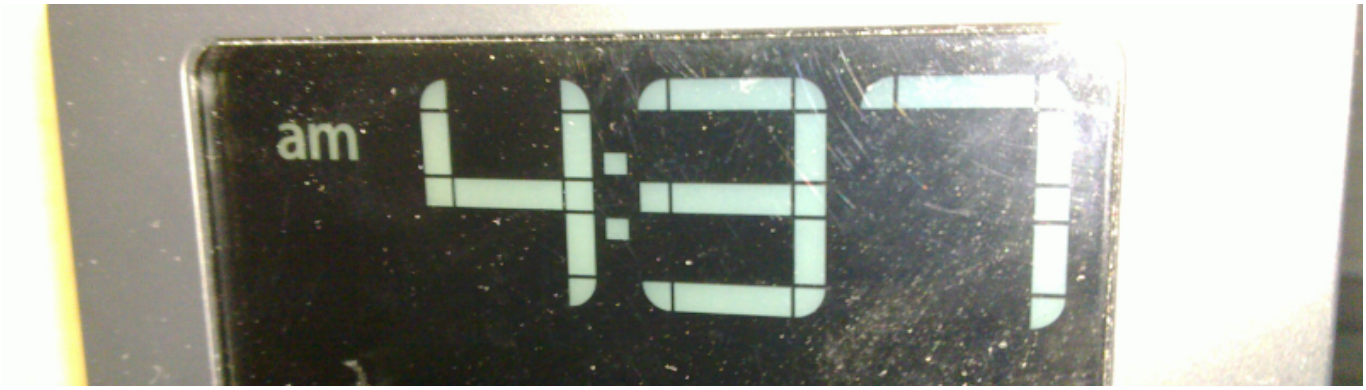
Pour obtenir la liste complète des présentations et les abstracts, il suffit de télécharger l'agenda de la conférence.

Pour résumer, la conférence était très “commerciale”, mais certaines présentations étaient néanmoins intéressantes :

- La plupart des conférences de la BioTeam : du workshop et de la conf. ;
- La présentation de Chris DWAN du Broad ;
- Et pour la culture générale, les présentations de Toby BLOOM et de Joseph SZUSTAKOWSKI.

Du côté des *buzz words* : **ScienceDMZ**, **Métadonnées** (mais personne n'a parlé de RDF) et **Consentement Patient**.

Jour 0 (Workshops)



Workshop 1 - "Aligning projects with Agile approch" par Gurpreet KANWAR

La présentation n'est pas disponible, mais plus de détails sont disponibles sur la page du workshop.

La plupart des participants viennent de “big pharma” et cherchent à passer à une méthode Agile.
La présentation commence par une vidéo sur youtube : Introduction to Scrum in under 10 minutes

Tout au long de la présentation Gurpreet mélange Agile, Scrum et XP.
Il s'avère que dans sa compagnie ils utilisent le plus souvent une approche “hybride” entre Waterfall et Agile pour obéir aux contraintes de la gestion de projet.
C'est cette partie PMO lourde (planification, cahier des charges...) que les participants cherchent à éliminer.

Parmi les slides de la présentation, certaines contiennent de bons résumés de :

- “why Agile” (#20) :
  - Time to market, frequency of release

  - Scope is not fully determined
  - Research project
  - No clear picture
  - Skilled developers are adaptable to change
  - Changing industry
  - Fully engaged client
- “Scrum roles” (#34) :
  - Product owner: identify features, prioritize them
  - The Team: create the product, self-manage
  - Scrum Master: escalation of blockers and risks, lead meetings, team building…
- “Project management steps” (#39)

Les points importants à retenir :

- Il faut être sûr d'avoir les bonnes personnes aux réunions, celles qui peuvent prendre des décisions, pas celles qui vont demander à leur supérieur.
  D'où l’importance d'avoir un bon agenda.
- Il faut rendre les problèmes visibles, car cela incite la personne en charge à les prendre en compte.
- Le team building est très important (“Pizza and donuts help!”)
- L'estimation du temps total est très importante
  Ils ont un outil (un excel) pour ça, et font appel à un cabinet externe

> Knowledge is knowing that a tomatoe is a fruit.
> Wisdom is knowing not to put it in a fruit salad!

Workshop 14 - "Converged IT infrastructure in life sciences" par l'équipe BioTeam

Ce workshop est en fait une mini-conférence faite par 4 consultants de la compagnie BioTeam.

0) "Introduction" par Ari BERMAN

Télécharger la présentation.

1) "Storage, HPC and Networking" par Aaron GARDNER

Télécharger la présentation.

Présentation assez moyenne. Les seules informations à retenir sont :

- GPFS est une bonne technologie
- La solution “ferme, cloud scientifique et cloud service” est une bonne solution

> Bring it all together to abstract the hardware from the user!

2) "Org, challenges, networks and Science DMZs, oh my!" par Ari BERMAN

Télécharger la présentation.

Excellente présentation sur les problématiques Bioinfo en général. On dirait du Sapay™.

Insiste sûr le fait que l'IT (UT6 dans notre cas) doit travailler avec les biologistes, car de toute façon ils auront toujours le dernier mot.
Il faut donc construire des solutions pour eux en fonction de cas d'utilisation concrets.
Le principal problème est que la bioinformatique change tous les 6 mois, alors qu'une infrastructure IT change tous les 2-5 ans.

Il souligne que le réseau, et pas la puissance de calcul ou le stockage, est en passe de devenir le point bloquant.
C'est pour cette raison que les ScienceDMZ ont été inventées.

3) "Converged infrastructure from cloud perspective" par Adam KROUT

Télécharger la présentation.

L'intervention la plus intéressante de la journée. C'est un tour d'horizon des technologies à regarder.

Pour Adam, “converged infrastructure” veut dire une infrastructure définie et manipulable par des logiciels, où tout a une API.

- Software defined network: amazon VPC
- Software defined storage: disk, database, object, block, data, filesystem and the life cycle policy management that goes with it
- Infrastructure as code: AWS cloudformation (CFNCluste on github), OpenStack Heat
- Configuration management: chef, puppets, ansible
- Software defined datacenter: Apache Mesos
- Lightning fast cluster computing: Apache Spark (good parallel programming primitives)
- Lambda architectures (lambda-architecture.net)
- Data flow frameworks: AWS Kinesis, AWS Lambda
- Build for parallel processing: multi-threaded, MPI

Ce que doit contenir une “Converged IT plateform”: resource management, automatic storage allocation, service discovery.
Tout doit être: distribué, monitoré (perfSONAR), logué (CloudWatch, CloudTrail) avoir une API, tourner dans un conteneur, avoir une politique de gestion de données.

Il faut envisager un datacenter comme un ordinateur : replication, pratitionning, load balancing, health checking, compression, consistency.

“Tool building vs Automation”: mieux vaut avoir une bonne boite à outils manuelle, qu'un système mal automatisé ou trop lourd à développé et maintenir.

Évoque Julia comme un langage à la fois flexible et performant.

> Don't build something that breaks all the time!

Souligne l'importance de créer des systèmes pour les humains :

- languages and frameworks do matter
- build the right abstractions

Insiste sur le fait qu'on doit apprendre en essayant et testant soit même ces solutions, “learn by doing”.

4) "The freedom to discover" par Bhanu REKEPALLI

Télécharger la présentation.

Bhanu part du constat que les super ordinateurs existants sont rarement utilisés pour les sciences de la vie.
Cela vient probablement du fait que les processus sont durs à paralléliser.
La suite de la présentation une suite sans fin de cas où les HPC ont été utilisés dans des projets de biologie.

Conclusion : le vrai gain de performance intervient quand on implémente l'algorithme in silico.

Jour 1

Le prix du petit déjeuner à l'hôtel est carrément prohibitif ! Heureusement, à Boston, on peut prendre un petit déjeuner équilibré pour à peine 2$!

"Precision Combination Therapy: Discover • Design • Deliver" par Chris SANDER de Memorial Sloan Kettering Cancer Center

> To elimintate cancer... you have to eliminate cigarettes!

Les nouvelles approches sont “combinées”. Le développement se fait en 3 étapes :

- Discover combination
- Design trials
- Deliver therapy

La première étape consiste à inférer, puis utiliser des modèles prédictifs de réseaux.
L'étude se fait suivant les étapes : perturb, measure, infer, predict.

Le problème pour la seconde étape est que tous les cancers sont différents. Il est donc nécessaire de créer des “sous groupes” de patients pour faire des tests valables.

Avant de passer à la présentation de quelques outils, Chris rappelle que les logiciels et les ordinateurs ne font pas tout !
Il faut des cerveaux pour réfléchir à la conception des logiciels et au design des tests.

Outils :

- cBioPortal for cancer genomics
- EVfold : prédiction de structure tertiaire de protéine en étudiant la coévolution des résidus qui ont des interactions.

"Winner of Benjamin Franklin Award" : Owen WHITE

Auteur d'un article paru dans Nature en 2003 : Unrestricted free access works and must continue.

Les données ne sont rien sans les métadonnées associées. Ce ne sont pas tant la technologie, le format et la transaction qui comptent, mais la sémantique !

Il faut avoir une communauté engagée qui utilise, recycle et fait évoluer les ontologies.

> I would like to thank my hot girlfriend that I'm marrying next month!

Best practices awards

- EPAM : Automated Accelerometric Detection of Epileptic Seizures in Rodents (“I hope our work has made the life of people and rodents better!”)
- Judge's price à la Mickael J Fox Foundation pour son utilisation de TranSMART

Session Posters

Beaucoup de stands de vendeurs, pas beaucoup de posters :

- Bioinformatics Brew (BiB) une sorte d'environment modules + gestionnaire de paquet pour la bioinformatique

"Trends from the trenches" par Chris de chez BioTeam

Encore une très bonne présentation technique de BioTeam.

Tous les sysadmin doivent apprendre à scripter, car tout aura bientôt une API.
Les ingénieurs réseau vont avoir besoin d'autre chose qu'une certification Cisco.
Univa GE est en très de dépasser OGE en terme d'adoption.
Les FGPA ne seront jamais vraiment utilisés, car ils sont trop durs à utiliser.
Les stockages objets sont l'avenir, principalement car ils permettent de prendre en charge facilement les métadonnées.
On peut faire des petastorage pas cher avec Linux, Lustre et ZFS.

Les défis à venir :

- l'espace de stockage ne sera pas toujours élastique
- les instruments de tous les domaines vont se mettre à produire, et plus uniquement les séquenceurs
- il deviendra encore plus important de partager les données

Il faut “go with the flow” et développer “décentralisé”. Le réseau va donc devenir le nouveau bottle neck… “prepare for pain” ! La connexion 10 Gbps va devenir le nouveau “normal”.

Chris reparle des Science DMZ, le buzz word de la conférence.

Note : Conférence Converged IT summit les 9-10 septembre à San Fransisco

"EVO:Rail" par Michael MCDONOUGH de VMWare

Présente la solution “data center in a box” de VMware.
Tout est inclus, matériel et licences, dans une appliance qui est up-and-running en 15 min. Bien sûr, les appliances (+/- 100 VMs) sont stackables.

"Power Gene #LetYourDataFly" par Franck LEE de chez IBM

Présentation presque embarrassante sous forme d'une répétitive métaphore aéronautique.
Jette des noms de technologies (TranSMART, Spark/ADAM), mais rien de bien concret.

"Security vs Freedom – it's not a matter of Philosophy!" par Nora MANSTEIN de chez Bayer

Télécharger la présentation.

(Je ne suis pas encore sûr d'avoir compris si l'oratrice était naïve ou cynique…)

Bayer veut se lancer dans la génomique, car ils sont en retard par rapport à la concurrence, et que cela leur permettrait de développer des best-sellers.
Ils achètent donc des données à analyser, mais voilà, il y a des lois à respecter quand on utilise des données patients… sinon on peut être condamné à des amendes.
Ils ont développé un outil moche et compliqué pour pouvoir gérer les échantillons utilisables en fonction des études.

"Privacy, Access Control and Security…" par Toby BLOOM du New York Genome Center

Très bonne présentation sur tous les problèmes rencontrés quand on veut faire des études à partir de données de patients.
On ne peut rien faire si les données génomiques ne sont pas associées aux données cliniques. Et même dans ce cas on a le problème de consentement trop restrictif (uniquement pour une maladie),

ce qui rend les données difficilement utilisables.
Le NIH est en train d'inciter à l'adoption du “consentement élargi” (et facilité) en ne subventionnant que ce type d'études.
Les données générées depuis une décennie restent un problème à régler.

"Managing Genomic Data at Scale!" par Jose L ALVAREZ de chez Seagate

Rien sur iRODS, juste du bla bla sur Seagate.

"Making Vizualisations Work" par Thimothy KROPP de la FDA

Télécharger la présentation.

En fait la présentation formalise plus la question que ce qu'elle n'apporte de réponse.

"Beyond parallel FileSystem" par James REANEY de chez SGI

Présentation technique des “disques” NVMe (consortium industriel) qui remplaceront bientôt tous les autres.
Le stockage est plus rapide et plus proche du processeur, il est meilleur pour les utilisations IO intensive et bandwidth intensive.

"The expanding face of metadata" par Steve WORTH de chez EMC²

S’il fallait encore le dire, Steve rappelle l'importance des métadonnées.
Celles-ci sous leur forme simple “key-value” ne sont pas encore adoptées, alors que les formes hiérarchiques arrivent.
Le coût du calcul des données baisse, mais le coût du stockage et du transfert ne va faire qu'augmenter.
De bonnes annotations peuvent diminuer ces coûts, car il devient plus simple de : partager, déplacer, sécuriser, vérifier, nettoyer, chercher…
Nous avons besoin d'“information scientists” qui maitriseront les ontologies.

Il présente iRODS comme une solution potentielle parmi beaucoup d'autres pour gérer ces annotations.
Ils sont en train de développer une interface d'administration et de gestion pour iRoDS : MetaLnx (?)

Jour 2

"How can non-professional be a driving force behind the medical breakthrougths of the future?" par Katherin WENDELSDORF de Qiagen

Présentation un peu commerciale pour encourager les patients à mettre leurs génomes sur Ingenuity et à la partager sur “Empowered Genome Community”. Les gens le faisant ont un accès gratuit à Ingenuity et peuvent partager leurs informations avec les chercheurs via l'outil. Lors des questions sur la “propriété des données”, elle évite soigneusement la question.

“Data is scarced, we need more well-phenotyped genomes.”

"-omics in the Era of patient-driven data collection" par Andreas KOGELNIK de l'Open Medecin Institute

De plus en plus de données viennent des “wearable”, avec plus de 14 paramètres monitorés en permanence.
Les nouvelles technologies permettent de recruter des patients mieux informés, plus vite.
Selon Andreas, la moitié de la population à une maladie chronique et “we all have a rare disease!”

"PatientsLikeMe" par Benjamin HEYWOOD

Les patients sont d'accord pour partager leurs données afin d'aider la science.
Grâce au site il est très facile de recruter des patients motivés.
Le problème des “wearables” et que les logiciels sont propriétaires et qu'on ne peut pas être sûr des informations qu'ils donnent. Les données sont elles significatives ?

"Intelligent infrascructure for Life Science" par George VACEK de DDN

Télécharger la présentation.

Présentation commerciale.

"How Next Generation Scale-Out Storage Fuels Sponsored by Breakthroughs

in Life Sciences" par Peter GODMAN de Qumulo

Fait la liste de toutes les fonctionnalités qu'un stockage “objet” devrait offrir, mais qui n'existent pas encore.

"Research Computing @Broad" par Chris DWAN du Broad Institute

Les trois points clés selon lui sont :

- Créer un environnement fédéré à base de cloud public
- Bien comprendre le cycle de vie de ses données
- Le consentement du patient, la confidentialité des données et la propriété de celles-ci sont les vrais problèmes

Ils ont mis en place OpenStack. Même s'ils n'en voient pas encore les bénéfices (car la manière de travailler n'a pas changé), ils ne regrettent pas du tout.
L'enregistrement des données dans un système de fichiers classique a ses limitations (path, links, metadata, nombre de fichiers...), mais les “vendors” n'apporteront pas la solution ! C'est aux “académiques” de la trouver, les vendeurs se chargeront de l'optimiser.
Pour Chris, les stockages objet sont pour les données “cool” et “safe”.

> Genome data is NOT de-identifiable

Les équipes de développeurs du Broad (une soixantaine de personnes) sont passées au développement Agile/Scrum.

"Start Small, Collaborate Often, Grow Big – Scaling Sponsored by NGS Compute and Storage Solutions for Personalized Medicine" par James Lowey du TGEN Institute

La présentation est principalement une pub pour la plateforme TGEN développée avec DELL.

"EMC² – Mixing Files & Objects" par Pactrick COMBES

Présentation commerciale.

"Breaking the $1,000 Genome Sequencing Barrier with Object Storage" par Andrew Crouse d'HudsonAlpha

Présentation sponsorisée par SwiftStack.

Commence par établir la liste des fonctionnalités à avoir : redundancy (geo), availability, runs on commodity hardware, no vendor lock-in, maintain staffing & expertise.
Les fonctionnalités intéressantes apportées : autodiscard (BCL files...), tmp URL, middleware to automatically run part of the workflow (BCL to FastQ...).

Il dit que leurs installations leur revient 10x moins chère que la même chose chez Amazon.

"Reproducible NGS Research: Practical Approaches and Case Studies" par Joseph SZUSTAKOWSKI de chez Novartis

Joseph décrit le processus d'audit annuel. Une figure de l'année passée, de la personne à auditer ou d'un de ses collègues, est choisie au hasard et il faut reproduire les résultats !

Conclusions : il faut éviter les solutions “point & clic” qui sont dures à documenter.
Dans l'équipe ils utilisaient knitr qui leur permettait de lancer les analyses et de documenter en même temps. Il recommande aussi la mise en place de checklists pour toutes les procédures afin d'être sûr que tout a été fait.

"The New tranSMART Platform v1.2 Provides Unparalleled Functionality for Translational Medicine" par Keith ELISTON, CEO de la TranSMART Foundation

TranSMART a été démarré pour trouver une solution au constat suivant : “50% des tests cliniques en phase 3 échouent”.

Pour tout le reste, il conseille d'aller voir leur chaine Youtube.