“Trends from the trenches”

Chris Dagdigian

2008 Bio-IT World Conference & Expo

Thank you for having me

- Infrastructure Track Speakers
  - 5+ Directors
  - 8+ Deans/Professor/Lab head
  - 6+ CEO/CTO/CIO
  - …
- Conclusion
  - I am one of the least accomplished speakers you will hear at this event

chris@bioteam.net

Thank you for having me

- Infrastructure Track Speakers
  - 5+ Directors
  - 8+ Deans/Professor/Lab head
  - 6+ CEO/CTO/CIO
  - …
- Conclusion
  - I am the least accomplished speaker you will hear at this event

chris@bioteam.net

Why I'm here

- The BioTeam
  - Scientists with production HPC skills
    - Bridging the gap between informatics & IT
  - Vendor & technology agnostic
- Often a resource for labs and workgroups that don't have their own supercomputing centers and IT empires
- In a given year …
  - Various levels of engagement with many clients
    - Gov/EDU/Biotech/Pharma/Fortune-20 clients
    - Work with lots of smart people on common problems
  - Result
    - Decent insight into how things operate "from the trenches"

chris@bioteam.net

Additional disclaimer

- Content of this talk may be inappropriate for some audience members
- Most BioTeam clients *don't* have 7 figure IT budgets, petabyte SANs and dedicated datacenters
- Will discuss problems that simply don't exist for the largest Bio-HPC centers
- Known bug:
  - I speak fast and carry a large deck

chris@bioteam.net

Topics (drawn from past 12 mo)

- Hardware
- Software
- Networking
- Storage & backup
- Data movement & workflow
- Neat stuff for ‘08

chris@bioteam.net

Observed Trends: Hardware

- CPU wars
  - AMD had the edge through 2006-2007
  - In 2008 we are back to benchmarking again

chris@bioteam.net

Observed Trends: Hardware

- Clever cooling
  - Not just massive wallmount HVAC …
  - I appreciate cooling systems that allow for diverse enclosures, rack and mounting strategies
  - What we've seen in the last year
    - Full APC "hot isle / cold isle" InfraStruXure enclosures
    - Standalone APC in-row (water) chillers
    - Liebert XDO overhead (compressed refrigerant) chillers

chris@bioteam.net

In-row chilling

1024 Core cluster @ Emory University

chris@bioteam.net

Sealed hot/cold isle enclosures

chris@bioteam.net

Liebert XDO Overhead Cooling

Site: Institute for Computational Biomedicine; Wiell Cornell Medical College

chris@bioteam.net

Liebert XDO @ Cornell (video)

chris@bioteam.net

Observed Trends: Hardware

- The small cluster market is mostly gone
  - Mostly talking about 2-8 node workgroup/lab clusters
  - Probably also taking the ‘desktop cluster’ market
- Replaced by SMP boxes with multi-core CPUs
  - 8 cores in 1U available from all vendors
  - ScaleMP: 16 cores & 128GB memory in 1 chassis
- Paraphrasing a colleague:
  - “*Server performance outpacing workflow requirements*”
  - “*The researcher who needed a small cluster last year now just needs a beefy workstation on her desktop*”

chris@bioteam.net

Observed trends: Power

- First seen in 2007:
  - Cluster nodes powered on and off automatically depending on size of the pending task queue
    - *Cornell Medical College, NYC*
  - Likely to become a more popular method
    - Coming: Grid Engine + Project Hedeby
      - SGE clusters that understand node power on|off

chris@bioteam.net

Observed Trends: Hardware

- Storage
  - Still have the same problems
  - Unhappy storage technology tradeoffs
    - The ‘exotic’ vendors offer blazing speed and a few features
    - The ‘mainstream’ vendors exclusively focused on enterprise
    - Both are really expensive
  - What I need
    - Massive scaling, decent speed & grab bag of enterprise features
      - Single namespace is ideal
      - Excellent management tools
      - Active Directory / AV integration
      - HIPPA and/or other compliance and audit features
      - Replication and sync features

chris@bioteam.net

Observed Trends: Storage

- Price spread on storage still extreme
  - Costco - 4/24/2008
    - 1TB external firewire drive for $229 USD
  - 4TB raw capacity: ~$2k to +$40k depending on vendor, features and performance
  - 100TB raw capacity: $200K to $1M
  - “Value” storage getting cheaper “faster”
    - Sun’s Thumper, Scalable’s JackRabbit, NexSan, etc.
    - 1TB SATA drives will extend this trend into 2008
  - “Enterprise” dropping far more slowly
    - Adding lots more features though …

chris@bioteam.net

Observed Trend: Storage

- Quoted costs for small cluster, all infrastructure and 100TB raw storage:
  - Same specs to multiple vendors
  - Vendor responses:
    - Commodity: $239,000 USD
    - Redundant commodity: $289,000 USD
    - Redundant + infiniband commodity: $392,000
    - Tier 1 / full integration: $948,000 USD

chris@bioteam.net

Observed Trends: Storage

- Continuing
  - Parallel & cluster filesystems becoming mainstream
  - Bloom off the rose for “HPC storage” market
  - Homework assignment:
    - Go back and check the website of any “HPC storage” vendor who contacted you in 2007
    - See how many have refocused on enterprise server virtualization;
      - Extra credit or dangerous drinking game:
        - Count references to “VMWare” on each homepage …

chris@bioteam.net

Observed Trends: Storage

- Storage Virtualization
  - Spent much of 2007 thinking this was just another dumb IT fad
  - Amen! Jacob Farmer made me see the light
    - Disk and spindle virtualization is very interesting
    - Still quite expensive; may change in 2008
      - Coolest innovators are still tiny companies
      - May need to wait on industry consolidation

chris@bioteam.net

Observed Trends: Backup

- My IT nightmare every year for the last decade
- 2007
  - Backup products not keeping up with daily advances in storage capacity promoted by vendors
    - Failing to keep up with both price and performance
- 2008
  - On it's way to becoming a sick joke
  - Storage products leave backup products in the dust
    - Almost too far ahead to even attempt to keep up

chris@bioteam.net

Observed Trends: Backup

- Everybody is on the D2D/VTL bandwagon
- Disk-to-disk is fine for some environments
- Some situations still require tape
  - All I can say about tape libraries, media cost and tape based solutions:
    - Aaarrggh!

chris@bioteam.net

Observed Trends: Backup

- My first encrypted tape deployment: 2007
  - 2008: Likely to encrypt almost everything
- Audience poll
  - How many here are encrypting backup media?
  - How many plan to in CY2008?
- Encrypted backups
  - A simple risk analysis makes this almost a requirement now

chris@bioteam.net

Observed Trends: Backup

- Encrypted backups (lessons learned)
  - Acquiring and implementing is not that hard
    - Personally I like wire speed hardware solutions that are independent of the backup software
      - Drive, library or appliance-driven
  - Certificate management difficult to do “right”
    - In 2007-2008 you will likely face this trade-off:
      - Data security vs. ease of operation/implementation
        - Convenient encryption == lower security
        - High security == hard to implement & manage

chris@bioteam.net

Observed Trends: Backup

- Encrypted backups (lessons learned)
  - Recommendation
    - Document specific requirement; then research solutions
    - My sole requirements:
      - Stay out of the newspaper
      - Mitigate legal risks of lost or stolen backup tapes only
    - My solution:
      - Chose the hardware-based solution that offered the simplest key management scheme yet was still friendly to use of complex “enterprise grade” procedures in the future

chris@bioteam.net

Observed Trends: Networking

- 10 Gigabit Ethernet is mainstream
- In 2007
  - Connect storage to networks
  - Connect switches together
- In 2008 …
  - Lots more switch to switch
  - Not sure about 10GB to server(s) adoption
  - Pricing no longer insane (see: arastra.com etc.)

chris@bioteam.net

Observed Trends: Networking

- Fast low-latency interconnects
  - Not much change since 2007
  - Infiniband seems slightly more mainstream now
  - Still low adoption rate in Bio HPC
    - Primary reason:
      - Parallel code quality & availability
  - People who are purchasing interconnects:
    - Often use them for filesystems rather than scientific applications

chris@bioteam.net

Observed Trends: Software

- Still blown away by the popularity of phylogeny applications
- Lots of demand in still for single purpose systems designed to run:
  - PAUP
  - MrBayes

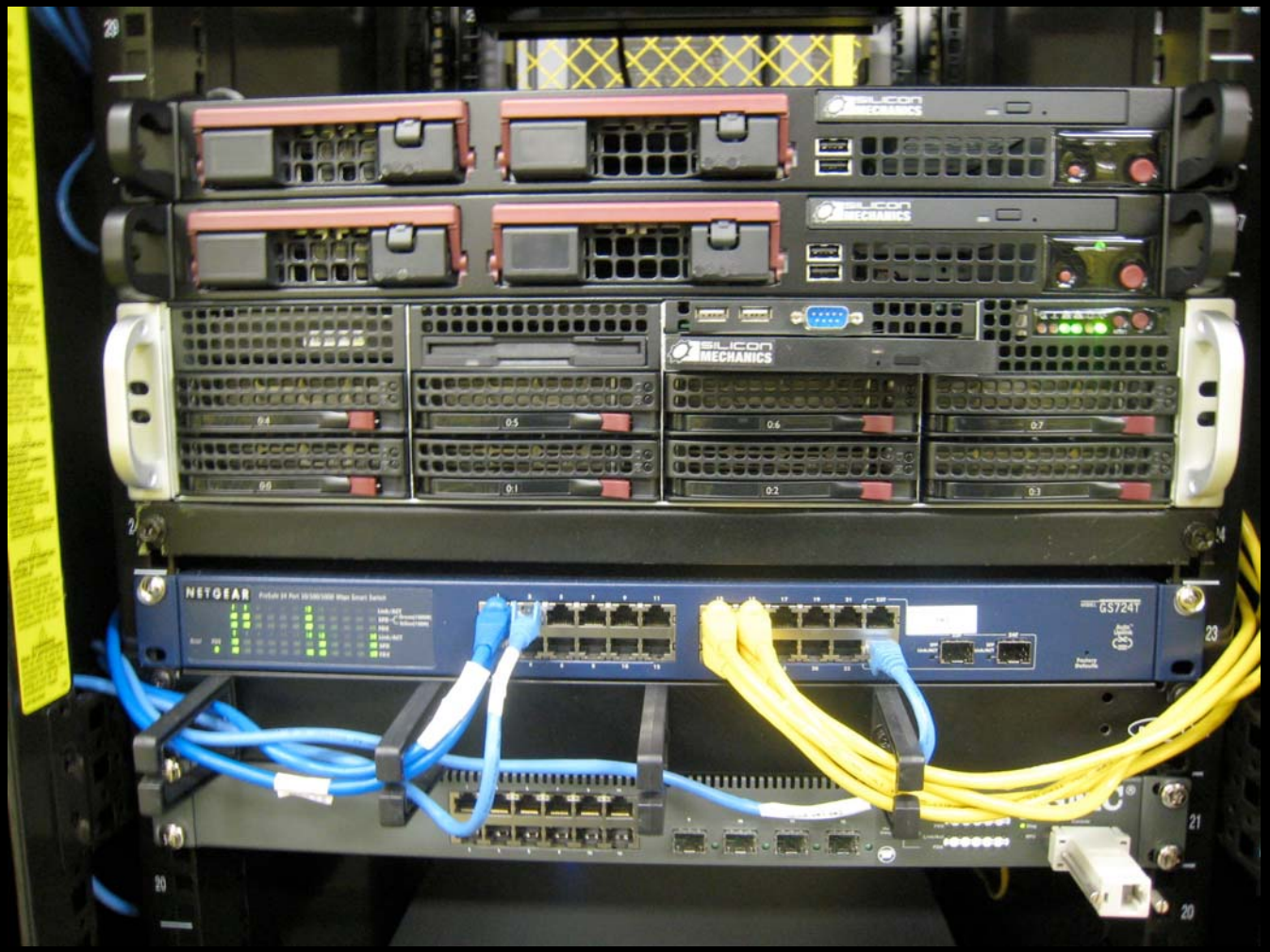
chris@bioteam.net

Emerging HW/SW Trend …

- Convergence of:
  - Quad-core 64bit CPUs available at all market segments
  - Very Large Memory
  - Fast disks
- … Plus
  - Commoditization of formerly high-end virtualization features
- Yields
  - Interesting deployment scenarios for virtualized platforms within the research datacenter

chris@bioteam.net

This kit ...



chris@bioteam.net

Replaces this …

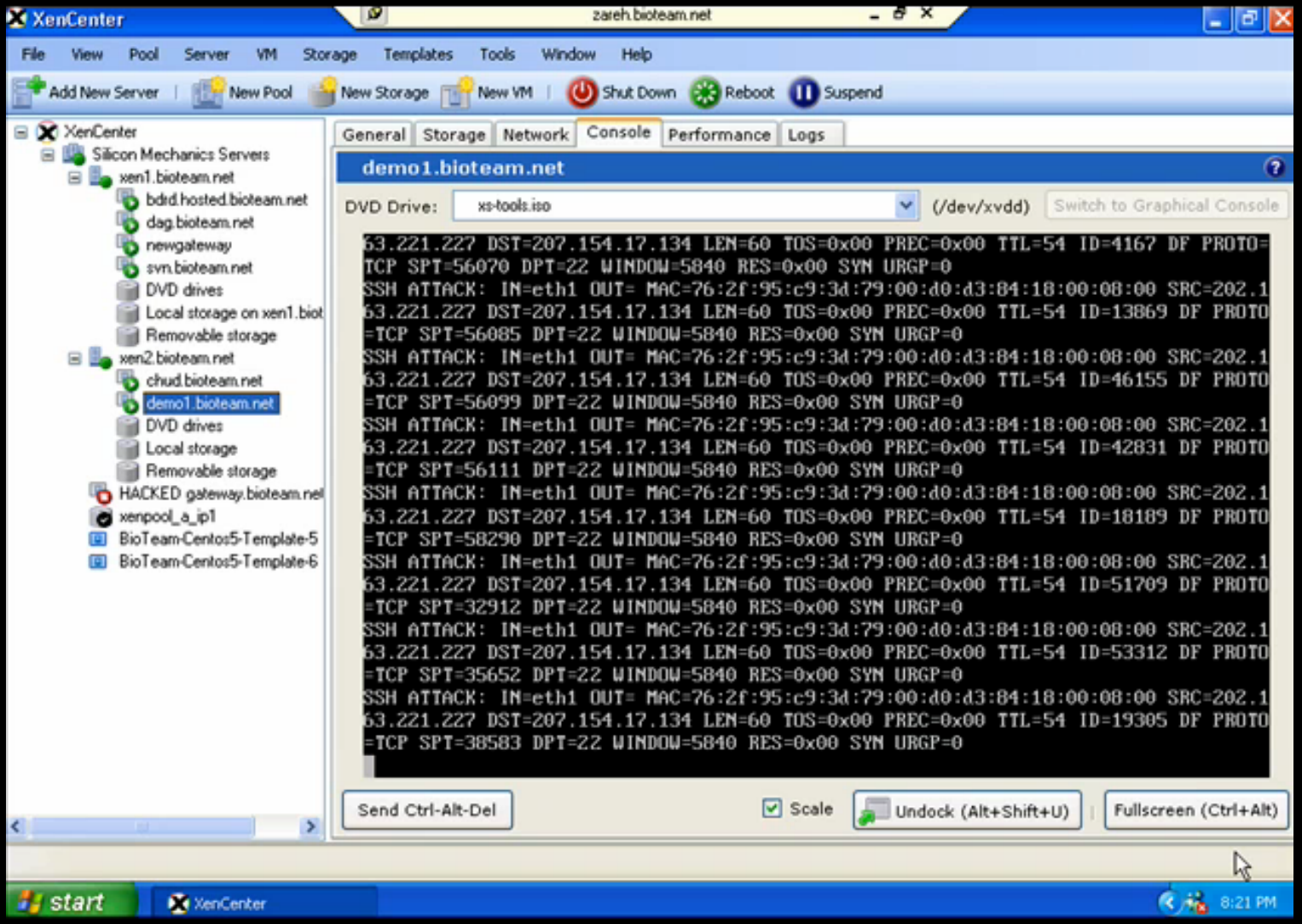
chris@bioteam.net

Virtualization is old news …

- The interesting trend is the rapid commoditization of “enterprise only” features
- Allows far more interesting usage
  - Client side & server side
- My favorite example
  - Live migration via vanilla NFS share

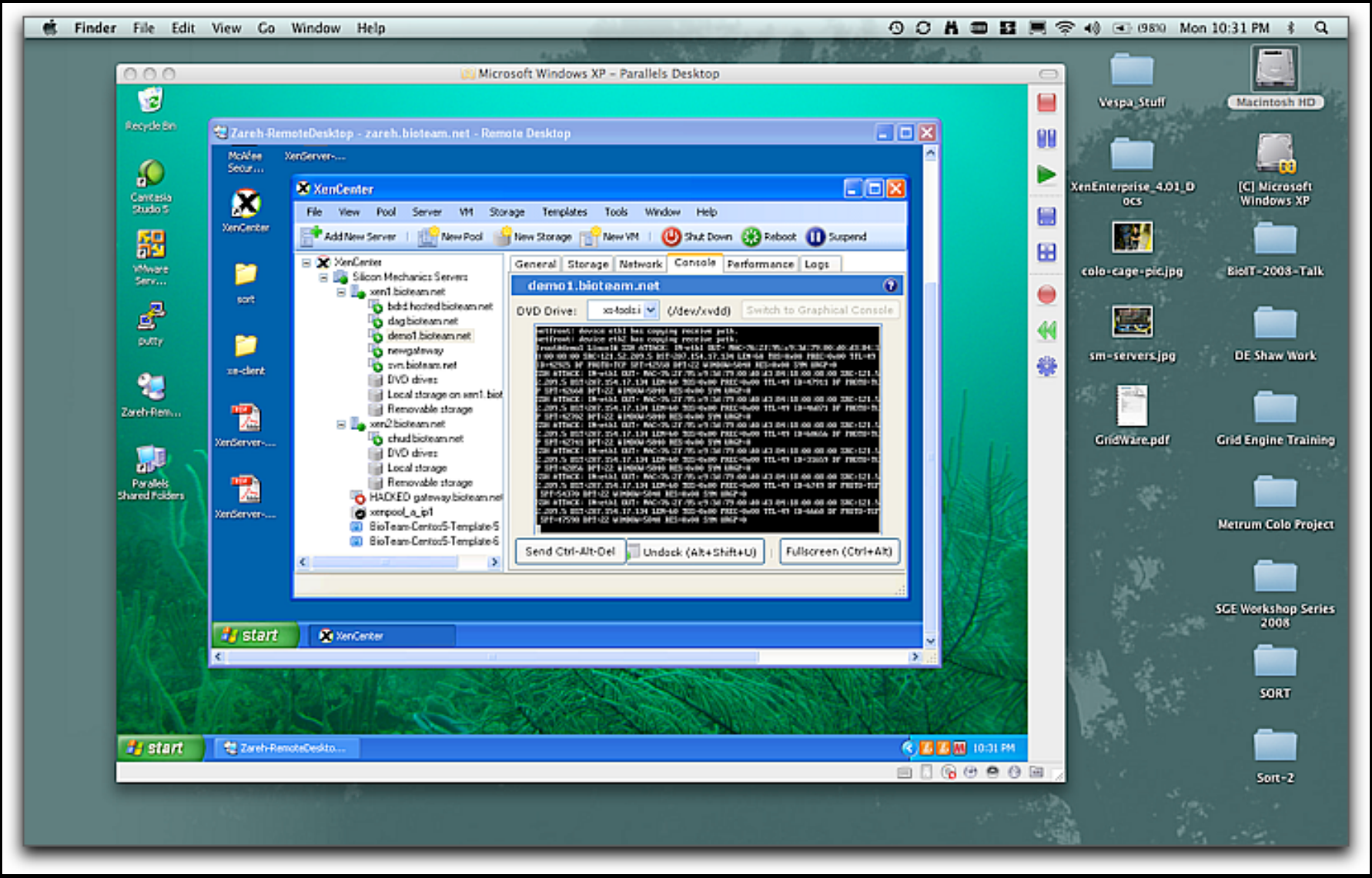
chris@bioteam.net

NFS-based Live Migration



chris@bioteam.net

Live migration: behind the scenes



chris@bioteam.net

Capital “G” GRID Computing

Capital “G” GRID Computing

- Remember the promise?
  - “Utility computing!”
  - “Like turning on a tap!”
  - “Multi-site? No problem!”
  - “Multi-entity? No problem!”
  - “Infinite capacity on demand!”
- GRID Facts (present):
  - Still a trainwreck for all but the showcase sites
  - At least the vendor FUD & empty press releases have died down
  - Only a tiny number of showpiece sites have the resources to do “GRID” computing for real

chris@bioteam.net

Capital “G” GRID Computing

- The GRID problem:
  - Compute power is dirt cheap and almost trivial to acquire
    - *Storage, backup and operation are the big costs*
  - “Bio IT” is often more I/O bound than CPU bound anyway
  - “Distribution” costs for research computing are high
    - Secure bandwidth between sites still expensive
      - *I/O bound? This is gonna get you …*
    - Holes punched in firewalls
    - Application integration difficulties
    - Certificate and keychain management complexity
    - Politics & policies ad nauseam

chris@bioteam.net

Capital “G” GRID Computing

- The GRID problem (continued):
  - The massive complexity and resources required for multi-site meta-scheduling
    - How many meta-scheduling experts does your org have?
    - How many holes or tunnels will be punched through your firewalls?
    - How to handle politics, user mapping and hostile sysadmins?
    - What local queuing & DRM features will you have to give up?
    - How many FTE’s will be required to keep it from falling over?
    - Who to blame *when* it falls over?

chris@bioteam.net

Capital “G” GRID Computing

- My $.02
  - Remember: I’m industry-centric and small-market focused
- Multi-site “GRID” computing does not (yet) deliver enough value to justify the time, expense, complexity and operational burden involved in building one.
  - Exceptions:
    - Non science drivers such as business continuity requirements or 24/7 workforce can help justify multi-site grid decisions
    - Companies like Univa (or “Univa UD” now)
      - … are on the right track

chris@bioteam.net

Capital “G” GRID Computing

- Please educate me
- I would love to learn about:
  - Any capital “G” GRID that
    - Spans multiple sites
    - Spans multiple institutional/political domains
    - Does something useful
    - AND:
      - Is not subsidized by funds from a national government
      - Is not subsidized by a commercial company trolling for reference sites, whitepapers and marketing quotes

chris@bioteam.net

The most terrifying trend …

What should be keeping you up at night

Terrifying trend: Terabyte Instruments

- 2007 was the tipping point
- We now have individual researchers with individual instruments that can:
  - *… generate terabyte scale data streams in a single experiment*
- Previously:
  - Terabyte data problems were at the workgroup, lab or organizational level

chris@bioteam.net

Terrifying: Terabyte Instruments

- The problem in a nutshell:
  - Individual researchers and/or single instruments are now capable of generating terabyte scale data *in a single experiment*.
    - Examples:
      - Confocal microscopy & Next generation DNA sequencers
  - These instruments are “cheap”
    - Easily affordable by grant-funded individuals and small labs
  - And …
    - Researchers don’t buy “just one” of these machines
    - Researchers may want to run them 24/7

chris@bioteam.net

Terrifying: Terabyte Instruments

- Why this is such a big deal
  - This is a nightmare even for the “big” centers with dedicated datacenters, large SANs and very competent IT staff
  - Imagine the effect on small organizations
    - The infrastructure and staff to support terabyte scale experimentation simply does not exist
  - Also
    - Researchers may be budgeting for the instrument and reagents but not the IT/operation requirements
    - Instrument vendors may be (intentionally or otherwise) downplaying the true infrastructure and operational costs of these instruments

chris@bioteam.net

Terrifying: Terabyte Instruments

- Is this your future?
  - Multi-terabyte storage resources in every wet lab?
  - *Sun Thumpers for all!*
- Tough decisions ahead
  - Centralized vs. decentralized data capture & movement
- This will effect *everyone* doing HPC “Bio IT”

chris@bioteam.net

Terrifying: Terabyte Instruments

- Central vs. Local storage
- In the last year we've done both
  - New construction (large R&D facility)
    - Confocal microscopy & advanced imaging
      - Storage: Centralized w/ lots of dark fiber to labs
      - Many VLANS within MPLS networking core
      - Redundant 10gig Ethernet to nearby closets
      - Can bring FC, iSCSI or Ethernet direct to instruments if needed
  - Existing facility with new DNA Sequencers
    - Storage: mostly local
      - Capture and 1st pass processing done locally
      - Replicate derived data to central store

chris@bioteam.net

Terrifying: Terabyte Instruments

- Audience poll (or find me afterwards)
  - Legit concern for your organization?
  - … or am I panicking over nothing?

chris@bioteam.net

One more (potential) trend …

chris@bioteam.net

Potential trend: Data Triage

- In 2007 we first saw
  - Deliberate decisions to not store primary data
- In the past
  - Always keep *all* data, essentially forever
    - Default excuses:
      - It costs to much to repeat the experiment
      - Experiment can't be repeated (imaging, microscopy)

chris@bioteam.net

Potential trend: Data Triage

- Moving forward (2008 and beyond)
  - Expect cost/benefit discussions among IT and scientific staff
    - Convey real costs of operating a research IT infrastructure
  - What data *really* needs to be kept?
    - Primary vs. Derived data
  - Given cost of storage+backup+operation costs …
    - In what cases is it actually be cheaper to rerun the experiment?
  - More info (articles & whitepapers):
    - http://blog.bioteam.net

chris@bioteam.net

Conclusion: Coolest 2008 trend

- AKA *"Future talk topics for next year …"*
  - Amazon Web Services
    - Amazon EC2 - Elastic Computing Cloud
    - Amazon S3 - Simple Storage Service
    - Amazon SQS - Simple Queue Service

chris@bioteam.net

Cloud computing w/ EC2

- Amazon's web services:
  - Utterly game changing
- I say this as …
  - A cynical hype-hating production-oriented corporate IT type

chris@bioteam.net

Amazon EC2

- Xen server instances on-demand
  - Starting at .10/hour for 32bit system
  - 64bit systems start at $.40/hour
  - Fire up as many as you need, whenever you need them
    - Many interfaces/control points
      - Mozilla plugins, CLI, Java, Perl, etc.

chris@bioteam.net

Amazon EC2

- Why it works
  - Smart pricing
    - Server instance pricing is reasonable
    - Traffic to/from S3 storage cloud is free
    - Experimenting is dirt cheap
      - 1 week of messing around == invoice for $9 USD
  - Easy to use
  - Clever people can make money
    - Amazon allows reselling AMI instance images
      - I can build a specialized workflow engine and charge a small fee on top of the Amazon costs
      - All financial transactions handled by Amazon
  - Limitations are pretty obvious
    - Pretty easy to know ahead of time what workflows are/are-not EC2 friendly

chris@bioteam.net

Amazon EC2

- Compelling economics
  - Consider: 100 CPU hour research problem
    - EC2: 10 large servers @ .40/hr for 10 hours
      - Work done in 10 hours at cost of $40 USD
    - EC2: 100 large servers @ .40/hr for 1 hour
      - Work done in 1 hour at a cost of $40 USD

chris@bioteam.net

Amazon S3

- Storage cloud
  - Popular with web 2.0 outfits
  - Required component of EC2 usage
    - All EC2 AMI (server images) are stored in S3
  - Cheap to move data in/out
  - Reasonable monthly fee for persistent storage
  - Free to move data within Amazon services

chris@bioteam.net

EC2 True Story

- Since start of 2008
  - Every single BioTeam consultant has independently deployed one or more EC2 solutions
    - No corporate mandate
    - Days, not weeks of development time
    - It just made sense
    - Satisfied many diverse use cases and deliverables

chris@bioteam.net

BioTeam & EC2 (since Jan 08)

- Individual apps in custom AMIs
  - Cross-platform python-based client GUIs
  - Mpiblast, mrbayes, etc.
- iNquiry product running within EC2
  - Scales to arbitrary size
- iNquiry data service moving to EC2/S3
- Grid Engine Training Clusters
  - Virtual classroom & training lab on demand
- Full blown Grid Engine clusters
  - Destined for production use
  - Spun up when needed
    - Per user, per-developer, per-workgroup

chris@bioteam.net

EC2 Limitations

- Personally not happy with 64 bit pricing
  - .40/hr is a big jump from the .10/hr 32 bit pricing
  - Would like a “small” 64 bit AMI instance type
- No promises on latency & location
  - AMI instances can be on different subnets
    - OpenMPI had issues with this …
  - Data movement of obvious concern
- Good news
  - Amazon adds features rapidly
    - Within last 1.5 months:
      - Elastic IP addresses
      - Availability zones (request US or European hosting)
      - Cheaper storage transfer rates
      - Support & service contracts

chris@bioteam.net

End;

- Thanks!
- Plug
  - http://gridengine.info
  - June ‘08 Grid Engine Workshops
- Questions?
- Comments/feedback:
  - “chris@bioteam.net”