PlanetLab: an open community testbed for Planetary-Scale Services

Timothy Roscoe et.al.

Wednesday, April 23, 2003

Intel Research Berkeley

PlanetLab today

- 130 nodes at 55 sites in 10 countries, 4 continents, …
- Universities, labs, Internet2, colo centers
- Active and growing research community
- Just beginning…

PLANETLAB

Intel Research Berkeley

Where did it come from?

- Sense of wonder
  - The next important thing in extreme networked systems
    - Post-cluster, post-Yahoo, post-Inktomi, post-Akamai, post-Gnutella, post-bubble?
- Sense of angst
  - NRC: “looking over the fence at networks”
    - Ossified internet (intellectually, infrastructure, system)
    - Next internet will emerge as overlay on current one (again)
    - Defined by its services, not its transport
- Sense of excitement
  - new class of services that spread over much of the web
    - CDN’s, P2P’s are the tip of the iceberg
  - architectural concepts emerging
    - scalable translation, dist. storage, dist. events, instrumentation, caching, management

Intel Research Berkeley

Missing: hands-on experience

- Researchers had no vehicle to try out their next *n* great ideas in this space

- Lots of simulations
- Lots of emulation on large clusters
- Lots of folks calling their 17 friends before the next deadline

- - but not the surprises and frustrations of experience at scale to drive innovation

Intel Research Berkeley

Confluence of Technologies

- **Cluster-based scalable distribution, remote execution, management, monitoring tools**
  - UCB Millennium, OSCAR, ..., Utah Emulab, ...
- **CDNS and P2Ps**
  - Gnutella, Kazaa, ...
- **Proxies routine**
- **Virtual machines & Sandboxing**
  - VMWare, Janos, Denali,... web-host slices (EnSim)
- **Overlay networks becoming ubiquitous**
  - xBone, RON, Detour... Akamai, Digital Island, ...
- **Service Composition Frameworks**
  - Yahoo, Ninja, .NET, WebSphere, etc.
- **Established internet 'crossroads' – colos**
- **Web Services / Utility Computing**
- **Authentication infrastructures**
- **Packet processing (layer 7 switches, NATs, firewalls)**
- **Internet instrumentation**

intel

PLANETLAB

Intel Research Berkeley

March '02 Underground Meeting

Washington
- Tom Anderson
- Steven Gribble
- David Wetherall

MIT
- Frans Kaashoek
- Hari Balakrishnan
- Robert Morris
- David Anderson

Berkeley
- Ion Stoica
- Joe Hellerstein
- Eric Brewer
- John Kubiatowicz
- Anthony Joseph
- Randy Katz

Intel Research
- David Culler
- Timothy Roscoe
- Gaetano Borriello
- Satya
- Milan Milenkovic
- David Tennenhouse

Duke
- Amin Vadat
- Jeff Chase

Princeton
- Larry Peterson
- Randy Wang
- Vivek Pai

Rice
- Peter Druschel

Utah
- Jay Lepreau

CMU
- Srini Seshan
- Hui Zhang

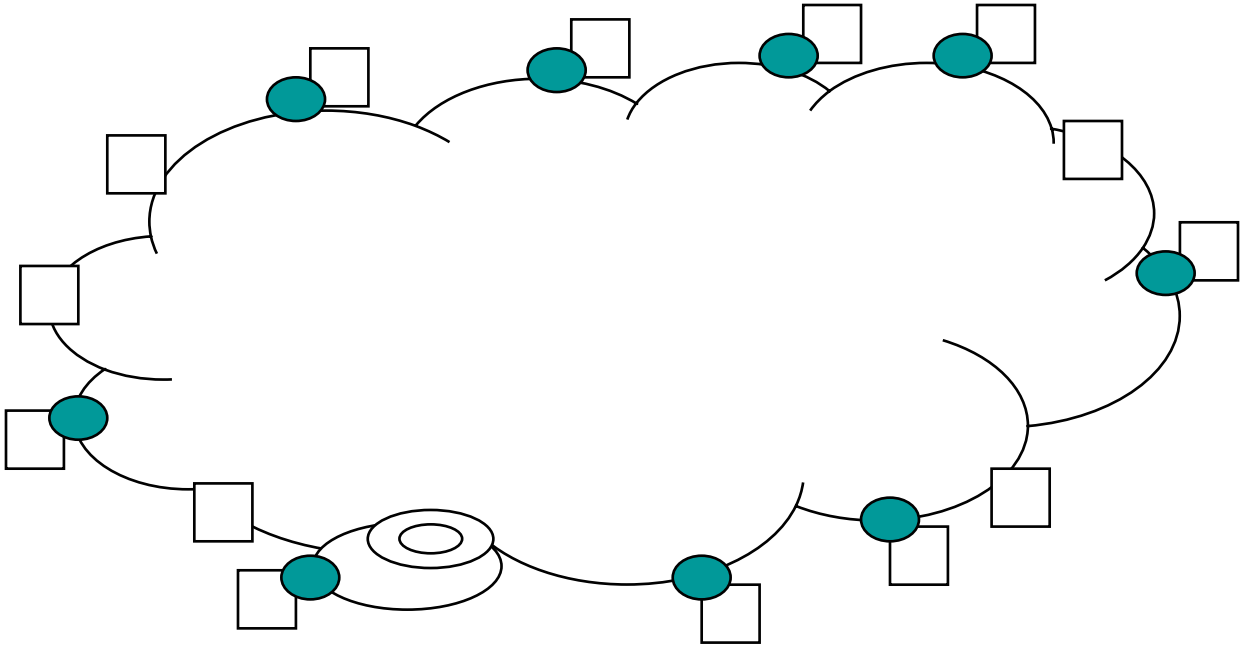
UCSD
- Stefan Savage

Columbia
- Andrew Campbell

ICIR
- Scott Shenker
- Eddie Kohler
- Sylvia Ratnasamy

Intel Research Berkeley

Guidelines (1)

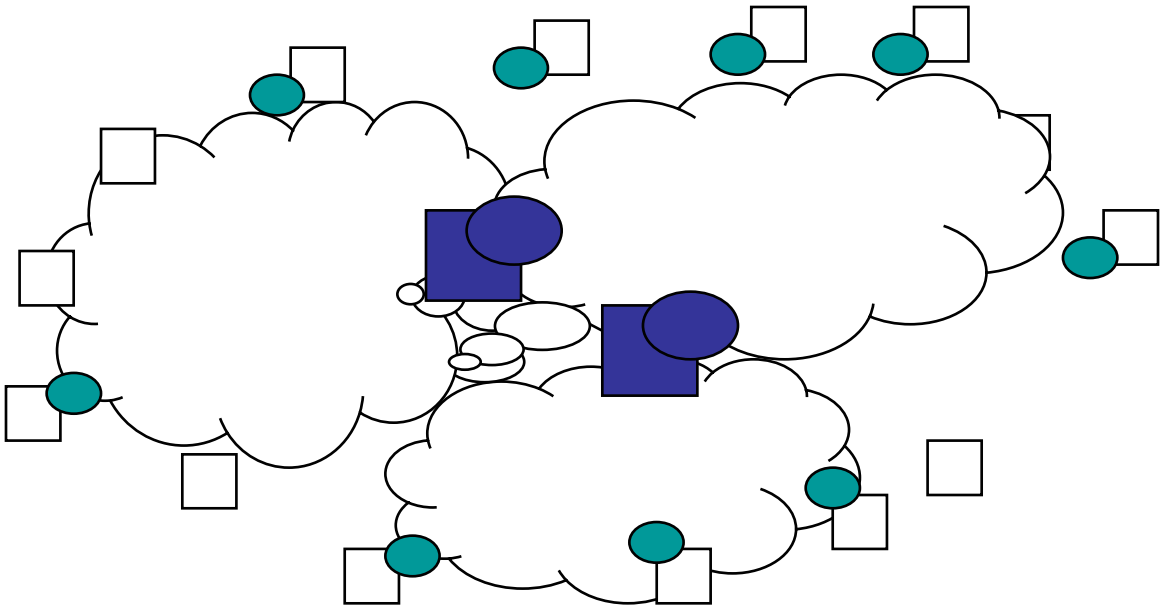


- Thousand viewpoints on “the cloud” is what matters
  - not the thousand servers
  - not the routers, per se
  - not the pipes

PLANETLAB

Intel Research Berkeley

Guidelines (2)



- you must have the vantage points of the crossroads
  - primarily co-location centers

Guidelines (3)

- Each service needs overlay covering many points
  - logically isolated
- Many concurrent services and applications
  - must be able to slice nodes ⇒ VM per service
  - service has a slice across large subset
- Must be able to run each service / app over long period to build meaningful workload
  - traffic capture/generator must be part of facility
- Consensus on “a node” more important than “which node”

Intel Research Berkeley

Guidelines (4)

- *Management, management, management*
- Test-lab as a whole must be up a lot
  - global remote administration and management
  - redundancy within
- Each service will require its own remote management capability
- Testlab nodes cannot “bring down” their site
  - generally not on main forwarding path
  - proxy path
  - must be able to extend overlay to user nodes?
- Relationship to firewalls and proxies is key

Intel Research Berkeley

Guidelines (5)

- Storage has to be a part of it
  - edge nodes have significant capacity
- Needs a basic well-managed capability
  - but growing to the seti@home model should be considered at some stage
  - may be essential for some services

Intel Research Berkeley

Outcome

- “Mirror of Dreams” project
- K.I.S.S.
  - Building Blocks, not solutions
  - no big standards, OGSA-like, meta-hyper-supercomputer
- Compromise
  - A basic working testbed in the hand is much better than “exactly my way” in the bush
- *“just give me a bunch of (virtual) machines spread around the planet,.. I’ll take it from there”*
- small distributed arch team, builders

Intel Research Berkeley

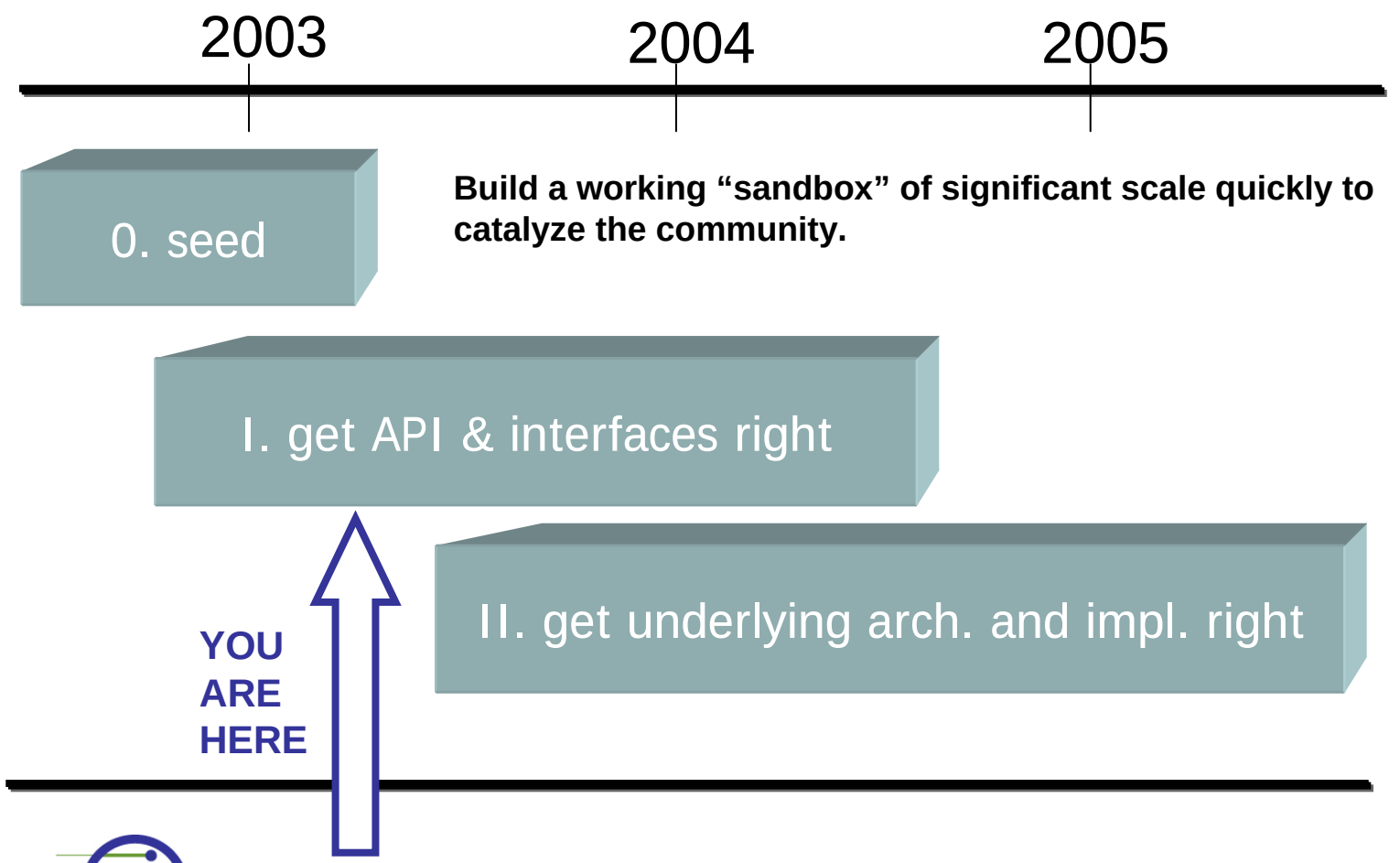
Tension of dual roles

- Research testbed
  - run fixed-scope experiments
  - large set of geographically distributed machines
  - diverse & realistic network conditions
- Deployment platform for novel services
  - run continuously
  - develop a user community that provides realistic workload

PLANETLAB

Intel Research Berkeley

Overlapping Phases



PLANETLAB

Intel Research Berkeley

Architectural principles

- *Slices* as fundamental resource unit
- Distributed Resource Control
- Unbundled Management
- Application-Centric Interfaces

- Self-obsolescence
  - everything we build should eventually be replaced by the community
  - initial centralized services only bootstrap distributed ones

Intel Research Berkeley

Slice-ability

- Each *service* runs in a *slice* of PlanetLab
  - distributed set of resources (network of VM)
  - allows services to run continuously
- VM monitor on each node enforces slices
  - limits fraction of node resources consumed
  - limits portion of name spaces consumed
- Challenges
  - global resource discovery
  - allocation and management
  - enforcing virtualization
  - security

Intel Research Berkeley

Unbundled Management

- Partition mgmt into orthogonal services
  - resource discovery
  - monitoring system health
  - topology management
  - manage user accounts and credentials
  - software distribution and updates
- Approach
  - management services run in their own slice
  - allow competing alternatives
  - engineer for innovation (minimal interfaces)

Intel Research Berkeley

Distributed Resource Control

- At least two interested parties
  - service producers (researchers)
    - decide how their services are deployed over available nodes
  - service consumers (users)
    - decide what services run on their nodes
- At least two contributing factors
  - fair slice allocation policy
    - both local and global components (see above)
  - knowledge about node state
    - freshest at the node itself

Intel Research Berkeley

Application-Centric Interfaces

- Inherent problems
  - stable platform versus research into platforms
  - writing applications for temporary testbeds
  - integrating testbeds with desktop machines
- Approach
  - take popular API (Linux), evolve implementation
  - later separate *isolation & application* interfaces
  - provide generic “shim” library for desktops

Intel Research Berkeley

Kick-off to catalyze community

- Seeded 100 machines in 42 sites July '02
  - avoid machine configuration issues
  - huge set of administrative concerns
- Intel Research, Development, and Operations
- UCB Rootstock build distribution tools
  - boot once from floppy to build local cluster
  - periodic and manual update with local modification
- UCB Ganglia remote monitoring facility
  - aggregate stats from each site, central database
- 10 Slices (accounts) per site on all machines
  - authenticate principal (PIs), delegation of access
  - key pairs stored in PL central, pushed out to nodes
- Basic SSH and scripts

Intel Research Berkeley

BootCD – enabling growth

- 2nd-Generation boot environment
  - Complete Linux distro on a CD
- Node *always* boots first from CD
  - Downloads signed script from bootsvr
  - Can fully install an OS
  - Can chain-boot a kernel
  - Can run remote secure diagnostics

Intel Research Berkeley

Service-Centric Virtualization

- VMs for complete desktop environment
  - e.g., VMware
  - extremely complete, poor scaling
- VM sandboxes widely used for web hosting
  - Ensim, BSD Jail, Linux VServers, UML,
  - limited /bin, no /dev, many VMs per ΦM
  - *limit the API for security*
- Scalable Isolation kernels (VMMs)
  - host multiple OS's on cleaner VM
  - Denali, Xen
  - Simple enough to make secure

Intel Research Berkeley

How much to virtualize?

- enough to deploy the next planet-lab within a slice on the current one...
- enough network access to build network gateways for overlays

- Phase 0: unix process as VM
  - SILK (Scout in Linux Kernel) to provide resource metering, allocation
- Phase 1: sandbox
  - evolved a constrained, secure API (subset)
- Phase 2: small isolation kernel with narrow API
  - some services built on it directly
  - host Linux / sandbox on top for legacy services

PLANETLAB

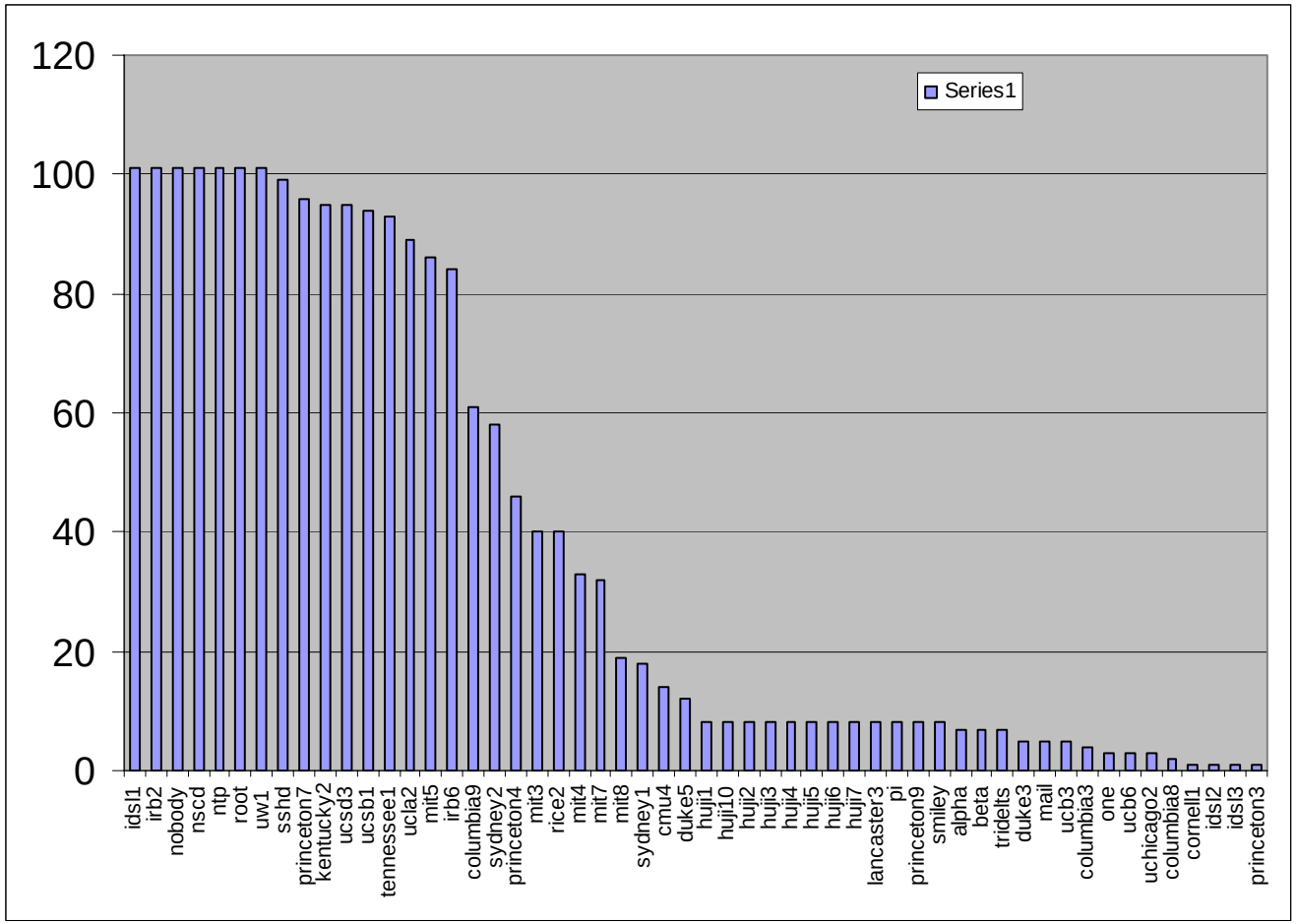
Intel Research Berkeley

VServer experience (Brent)

- New set of scaling issues: disk footprint
- Implemented VM-specific copy-on-write
  - O(1000) VMs per disk
  - Currently 200+ per node
- VMs are *cached* to reduce creation time (2-3 seconds)
- Slice login -> VServer root
- Limitations
  - common OS for all VMs (little call for multiple OS's)
  - user-level NFS mount
  - incomplete self-virtualization
  - incomplete resource isolation (eg. buffer cache)
  - imperfect (but so far unbroken) kernel security
- Raised bar for Isolation Kernels
  - May end up only as mechanism for multiple OSes

Intel Research Berkeley

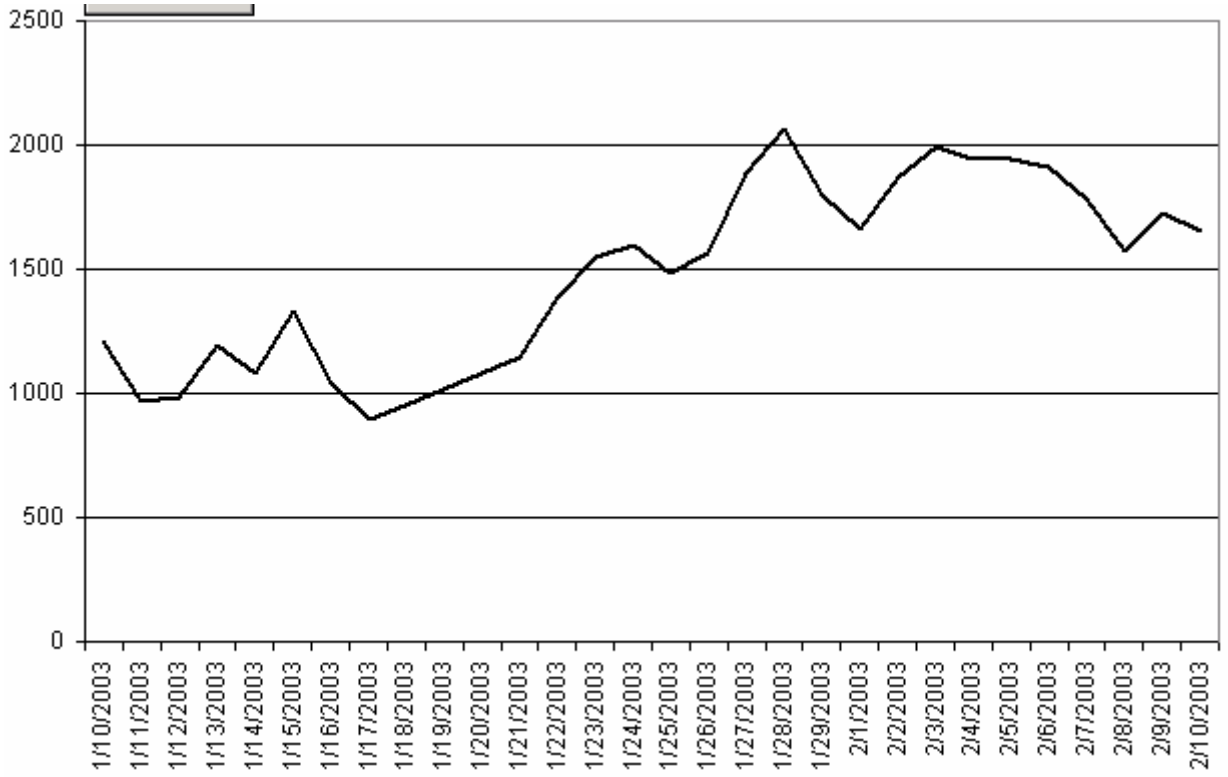
A typical day



PLANETLAB

Intel Research Berkeley

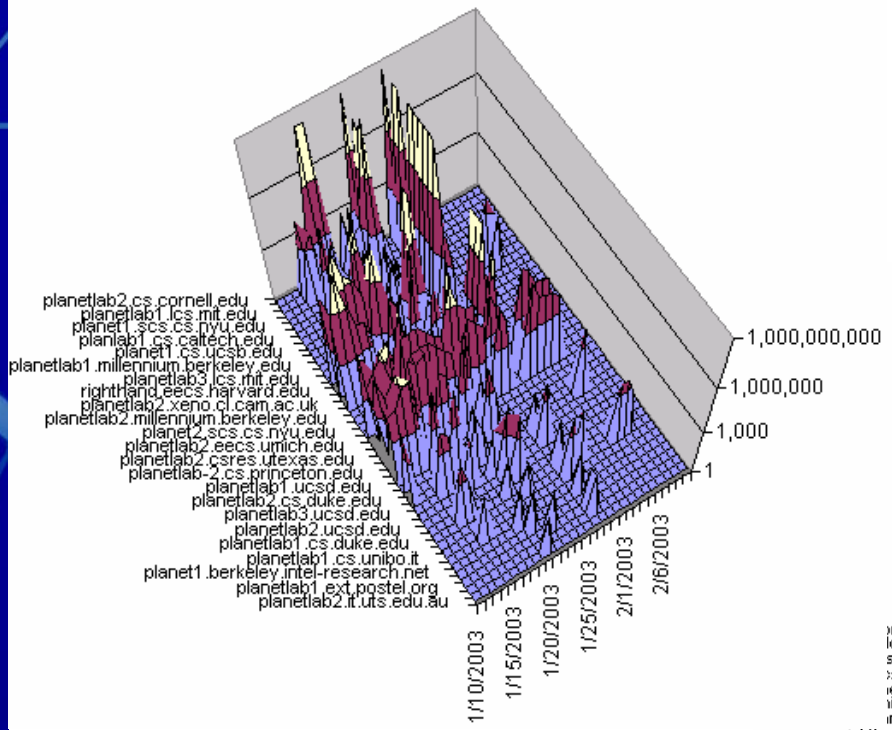
Pre-SIGCOMM deadline

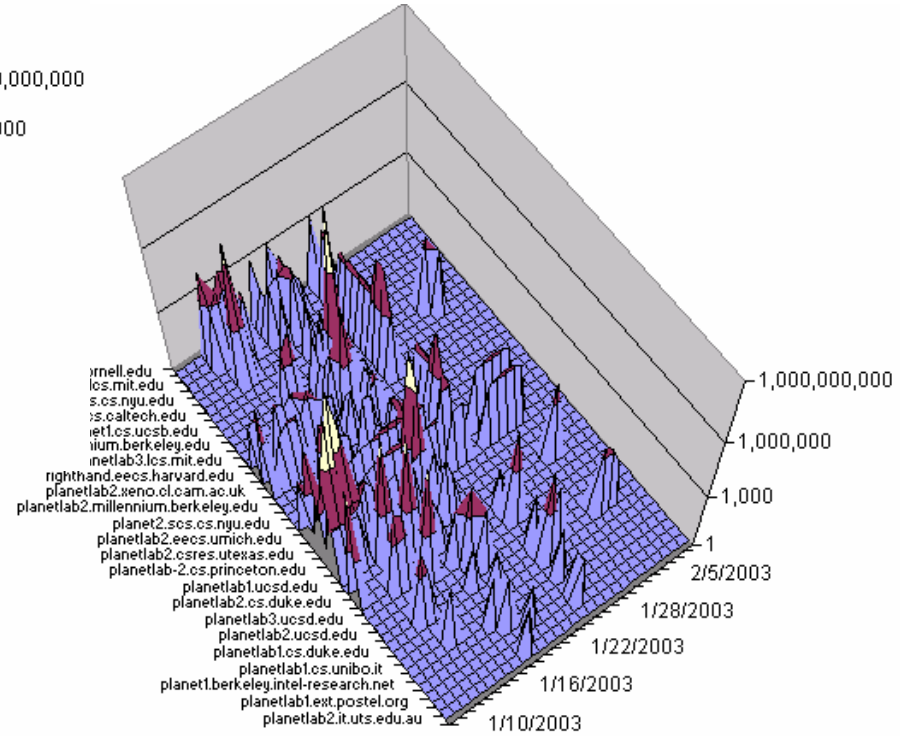


PLANETLAB

Intel Research Berkeley

A Slice for a Month (Duke)





PLANETLAB

Intel Research Berkeley

So what are people running?

intel

PLANETLAB

Intel Research Berkeley

No, really…

- Network measurement
  - Scriptroute, PlanetProbe, I3, etc.
- Application-level multicast
  - ESM, Scribe, TACT, etc.
- Distributed Hash Tables
  - Chord, Tapestry, Pastry, Bamboo, etc.
- Wide-area distributed storage
  - Oceanstore, SFS, CFS, Palimpsest, IBP
- Resource allocation
  - Sharp, Slices, XenoCorp, Automated contracts
- Distributed query processing
  - PIER, IrisLog, Sophia, etc.
- Content Dist. Networks
  - CoDeeN, ESM, UltraPeer emulation, Gnutella mapping
- Management and Monitoring
  - Ganglia, InfoSpect, Scout Monitor, BGP Sensors, etc.
- Overlay Networks
  - RON, ROM++, ESM, XBone, ABone, etc.
- Virtualization and Isolation
  - Xen, Denali, VServers, SILK, Mgmt VMs, etc.
- Router Design implications
  - NetBind, Scout, NewArch, Icarus, etc.
- Testbed Federation
  - NetBed, RON, XenoServers
- Etc., etc., etc.

Intel Research Berkeley

Ossified or fragile?

- One group forgot to turn off an experiment
  - after 2 weeks of router being pinged every 2 seconds, ISP contacted ISI and threatened to shut them down.
- One group failed to initialize destination address and ports (and had many virtual nodes on each of many physical nodes)
  - worked OK when tested on a LAN
  - trashed flow-caches in routers
  - probably generated a lot of unreachable destination traffic
  - triggered port-scan alarms at ISPs (port 0)
  - n^2 probe packets trigger other alarms

PLANETLAB

Intel Research Berkeley

The Gaetano advice

- for this to be successful, it will need the support of network and system administrators at all the sites...
- it would be good to start by building tools that made their job easier

Intel Research Berkeley

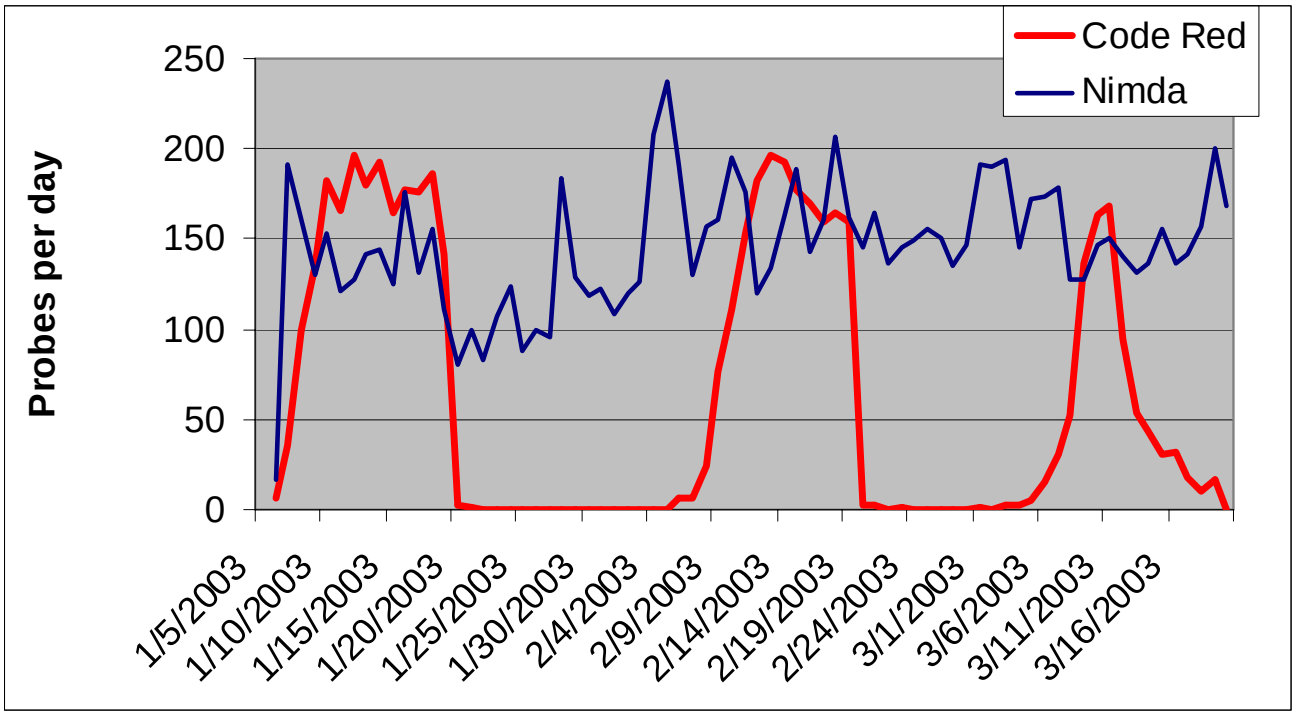
NetBait serendipity

- Brent deployed a simple webserver on each node to explain what PlanetLab was about
- It also logged requests…
- Sitting just outside the firewall of ~40 universities...
- A very large *honey pot*
- Shocking number of worm probes from compromised machines
- Imagine the epidemiology

PLANETLAB

Intel Research Berkeley

One example

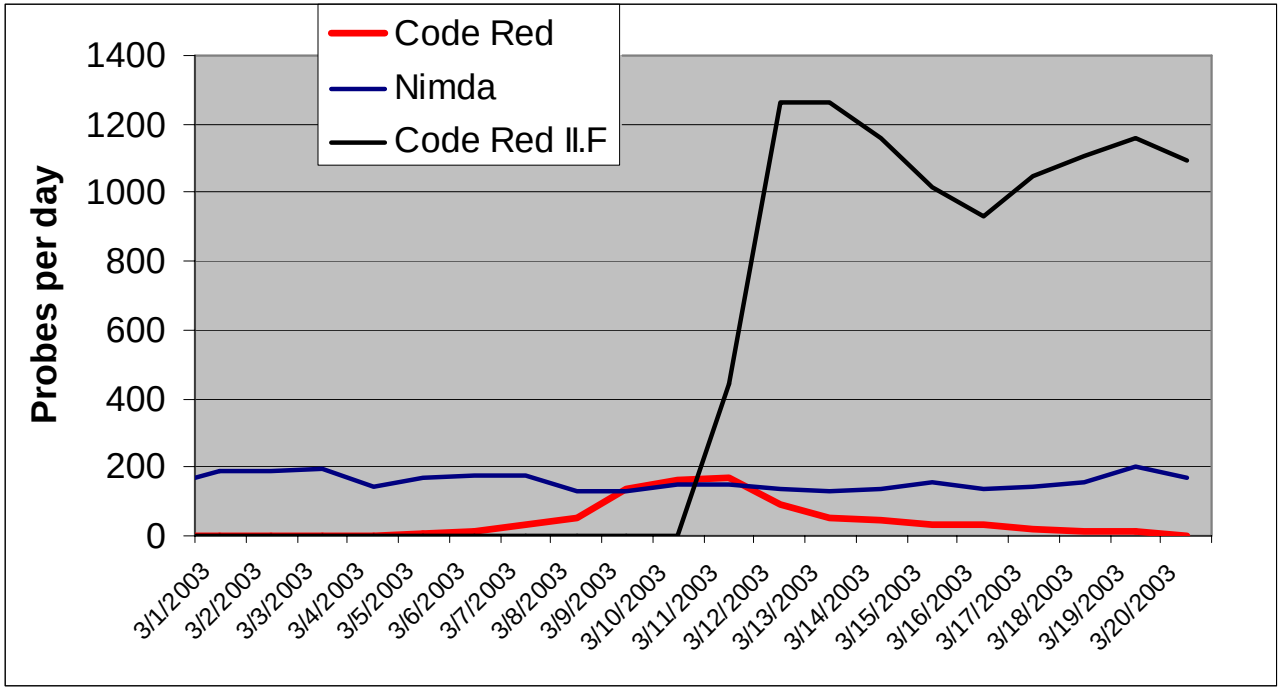


- The monthly Code Red cycle in the large
- What happened mid-March?

PLANETLAB

Intel Research Berkeley

No, not Iraq…

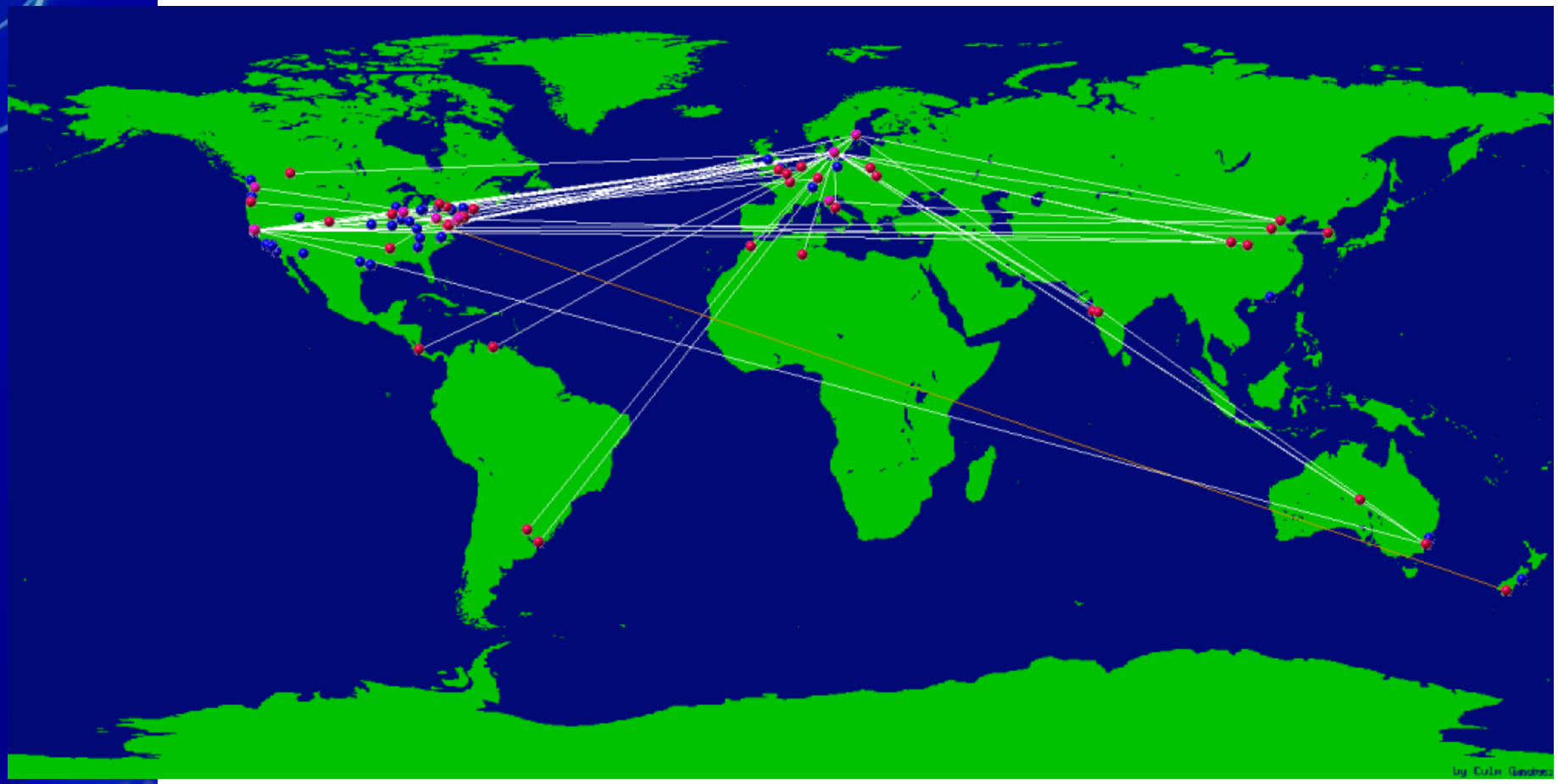


- A worm appeared and displaced the older Code Red

PLANETLAB

Intel Research Berkeley

Netbait view of March



intel

PLANETLAB

Intel Research Berkeley

What PlanetLab is about

- **Create the open infrastructure for invention of the next generation of wide-area ("planetary scale") services**
- **The foundation on which the next Internet can emerge**
  - Think beyond TCP/UDP/IP/DNS/BGP/OSPF…
  - …as to what the net *provides*
  - building-blocks upon which services will be based
  - "the next internet will be created as an overlay on the current one"
- **A different kind of network testbed**
  - not a collection of pipes and giga-pops
  - not a distributed supercomputer
  - geographically distributed network services
  - alternative network architectures and protocols
- **Focus and Mobilize the Network / Systems Research Community to define the emerging internet**

PLANETLAB

Intel Research Berkeley

Where is it going?

- It is just beginning
  - towards representative sample of the Internet
- Working Groups
  - Virtualization, Dynamic slices, Monitoring, etc.
- Building the consortium
  - Industrial partners, gov't funding, etc.
- Hands-on experience with wide-area services at scale is mothering tremendous innovation
  - nothing “just works” in the wide-area at scale
- Rich set of research challenges ahead

PLANETLAB

Intel Research Berkeley

May 8th FTF Meeting

- "Planetary-scale Services"
- Focus on:
  - Application research agenda
  - Infrastructure research agenda
  - Vision of the Planetary Services world
- Participation from IT eagerly sought

Intel Research Berkeley