An Integrated Experimental Environment for Distributed Systems and Networks

B. White, J. Lepreau, L. Stoller, R. Ricci, S. Guruprasad,
M. Newbold, M. Hibler, C. Barb, A. Joglekar

University of Utah

www.netbed.org

December 10, 2002

A Need for Diverse Approaches

- Simulation
  - Presents controlled, repeatable environment
  - Loses accuracy due to abstraction
  - e.g., ns, GloMoSim, *x*-sim [Brakmo'96]
- Live-network experimentation
  - Achieves realism
  - Surrenders repeatability
  - e.g., MIT "RON" testbed, PlanetLab
- Emulation
  - Introduces controlled packet loss and delay
  - Requires tedious manual configuration
  - e.g., Dummynet, nse [Fall'99], Trace Modulation [Noble'97], ModelNet [Vahdat'02]

Netbed

- Integrated access to:
  - Emulated, …
    - Allocated from a dedicated cluster
  - Simulated, …
  - Wide-area nodes and links
    - Selected from ~40 geographically-distributed nodes at ~30 sites
- Universal, remote access: 365 users
- 2176 "experiments" in 12 month period
- Time- and space-shared platform
- Enables qualitatively new research methods in networks, OSes, distributed systems, smart storage, …

Key Ideas

- "Emulab Classic"
  - Brings simulation's efficiency and automation to emulation
  - 2 orders of magnitude improvement in configuration time over a manual approach
- Virtual machine for network experimentation
  - Lifecycle & process analogy
  - Integrates simulation, emulation, and live-network experimentation

Two Emulation Goals

1. Accurate:
   Provide *artifact-free* environment
2. Universal:
   Run *arbitrary* workload: any OS, any code on "routers", any program, for any user

- Therefore, our default resource allocation policy is *conservative:*
  - Allocate full real node and link: no multiplexing
  - Assume maximum possible traffic

A Virtual Machine for Network Experimentation

Maps common abstractions …	To diverse mechanisms
Nodes	**Cluster nodes, VMs on wide-area nodes, ns**
Links	**VLANs, tunnels, Internet paths**
Addresses	IPv4, ns node identifiers
Events	**distributed event system, ns event system**
Program Objects	remote execution, ns applications
Queuing Disciplines	on simulated and emulated nodes
Projects, Users, Experiments	Independent of experimental technique
Topology Generation	Configure real or simulated nodes
Topology Visualization	View hybrid topologies
Traffic Generation	ns models, TG

Netbed Virtual Machine

- Achieved through OS techniques:
  - Virtualization/abstraction
  - Single namespace
  - Conservative resource allocation, scheduling, preemption
  - Hard/soft state management

- Benefits:
  - Facilitates interaction, comparison, and validation
  - Leverages existing tools (e.g., traffic generation)
  - Brings capabilities of one technique to another (e.g., nse emulation of wireless links)

Outline

- Background and Related Work
- **Experiment Life Cycle**
- Efficiency and Utilization
- New Experimental Techniques
- Summary

Experiment

- Acts as central operational entity
- Represents …
  - Network configuration, including nodes and links
  - Node state, including OS images
  - Database entries, including event lists
- Lasts minutes to days, to weeks, to … forever!

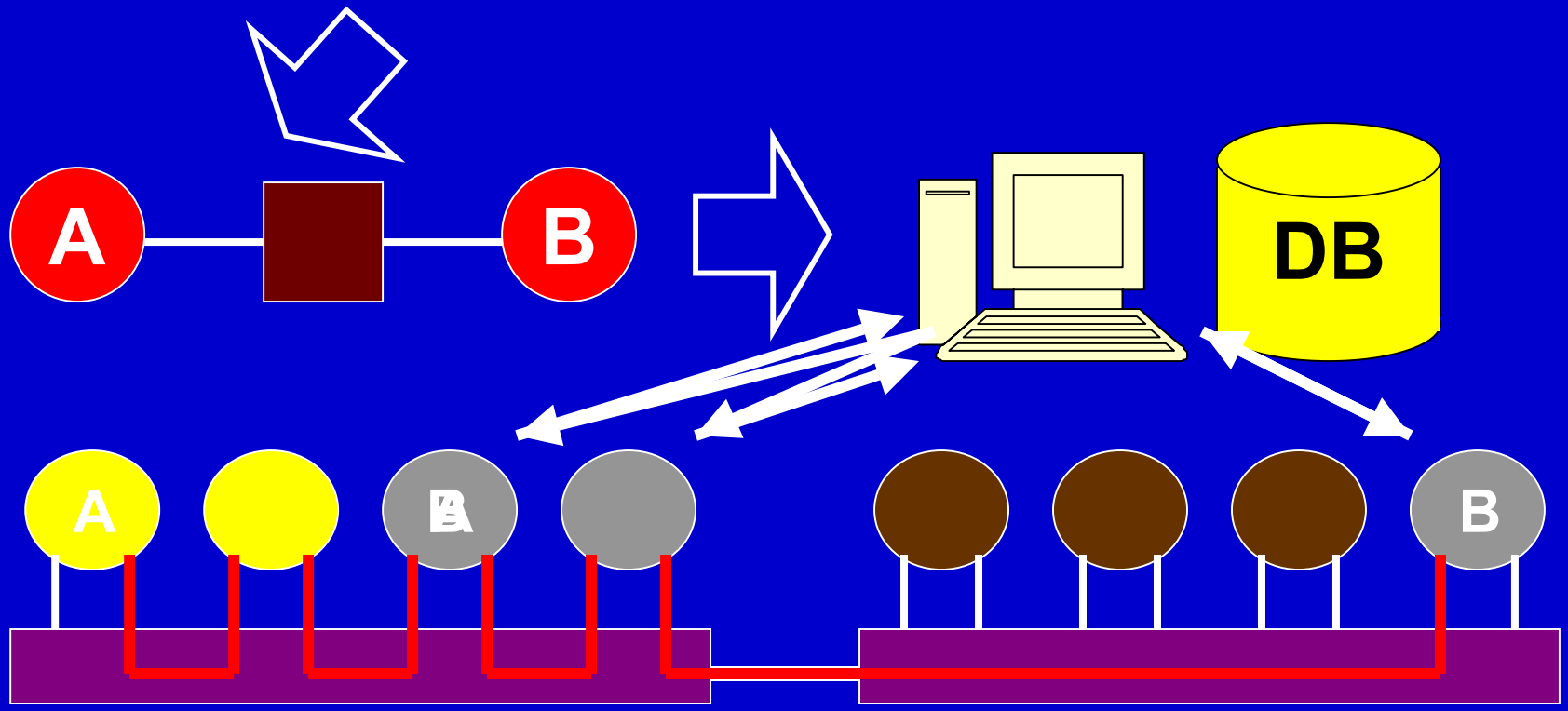
Experiment Life Cycle

- Specification
- Parsing
- Global resource allocation
- Node self-configuration
- Experiment control
- Preemption and swapping

Experiment Life Cycle

Gl[illegible]

```
$ns duplex-link $A $B 1.5Mbps 20ms
```



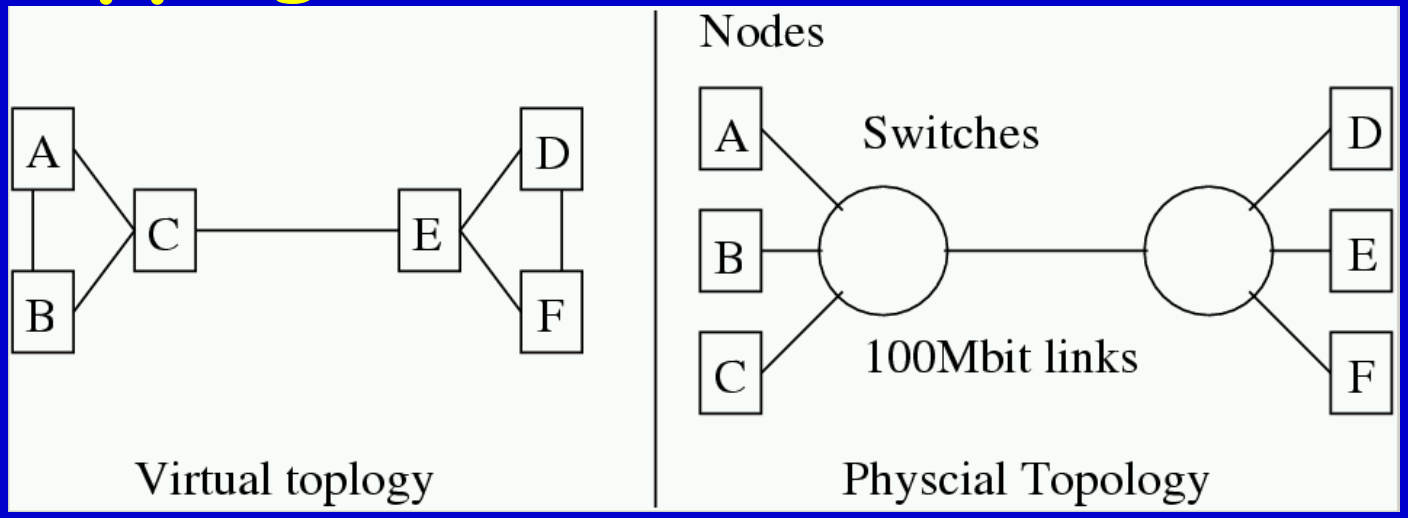
ns Specification

- ns: de-facto standard in network simulation, built on Tcl
- Important features:
  - Graceful transition for ns users
  - Power of general-purpose programming language
- Other means of specification:
  - Java GUI
  - Standard topology generators

Outline

- Background and Related Work
- Experiment Life Cycle
- **Efficiency and Utilization**
- New Experimental Techniques
- Summary

assign: Mapping Local Cluster Resources

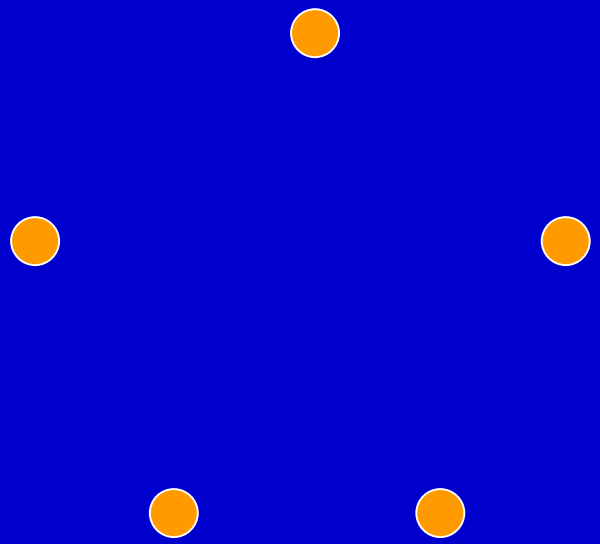


- Maps virtual resources to local nodes and VLANs
- General combinatorial optimization approach to NP-complete problem
- Based on simulated annealing
- Minimizes inter-switch links & number of switches & other constraints …
- All experiments mapped in less than 3 secs [100 nodes]

wanassign: Mapping Distributed Resources

- Constrained differently than local mapping:
  - Treats physical nodes as fully-connected (by Internet)
  - Characterizes node types by "last-mile" link
- Implements a genetic algorithm

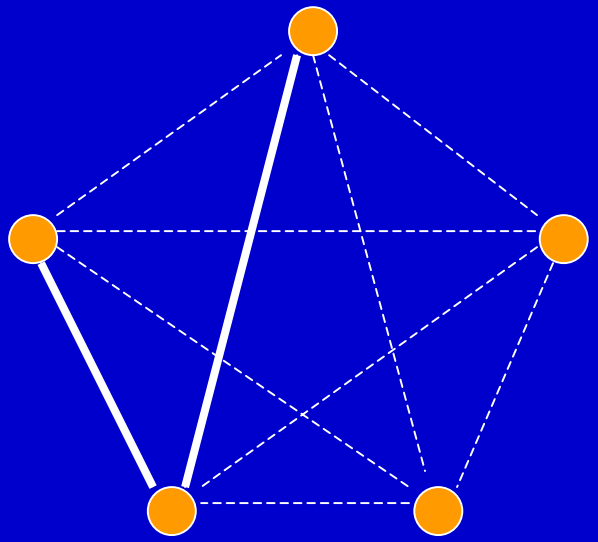
Mapping by Node Type



```
set src    [$ns node]
set router [$ns node]
set dest   [$ns node]

tb-set-hardware $src pc-internet
tb-set-hardware $router pc-internet2
tb-set-hardware $dest pc-cable
```

Mapping by Link Characteristics



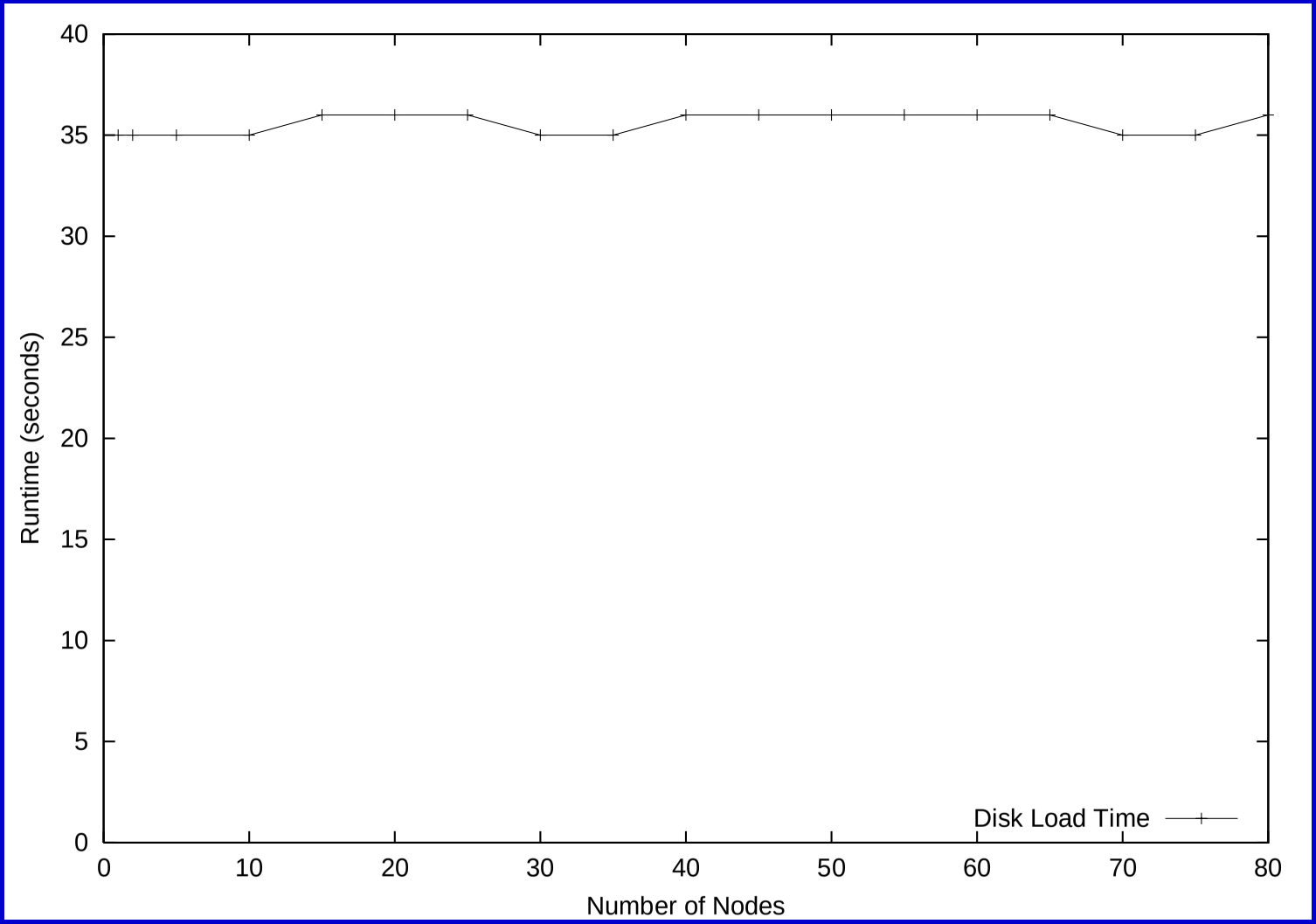
```
set src    [$ns node]
set router [$ns node]
set dest   [$ns node]

$ns duplex-link $src $router 10Mb 20ms DropTail
$ns duplex-link $router $dest 5Mb 100ms DropTail
```

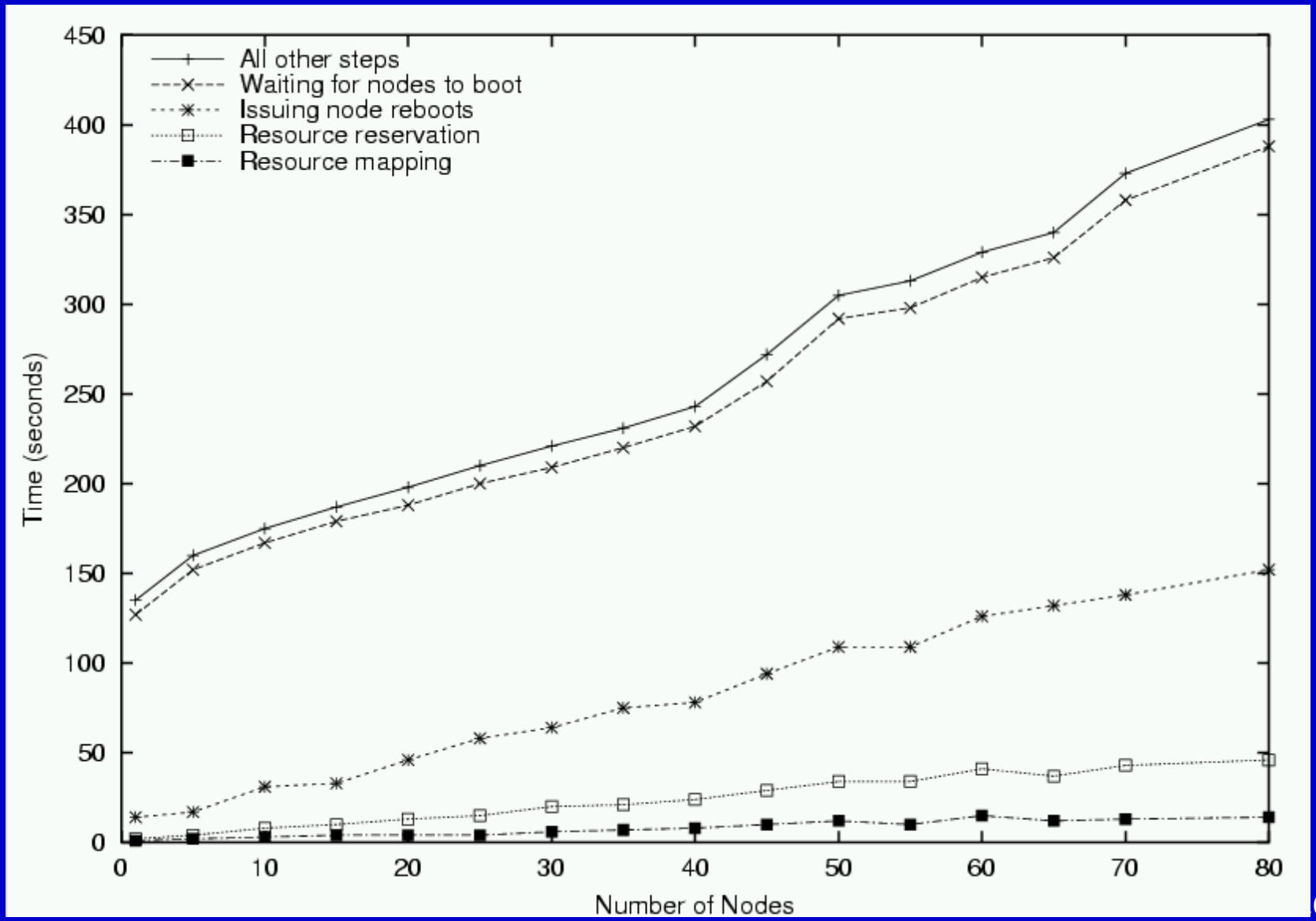
Disk Loading

- Loads full disk images
- Performance techniques:
  - Overlaps block decompression and device I/O
  - Uses a domain-specific algorithm to skip unused blocks
  - Delivers images via a custom reliable multicast protocol

"Frisbee" Disk Loader Scaling



Experiment Creation Scaling



Configuration Efficiency

- Emulation experiment configuration
  - Compared to manual approach using a 6-node "dumbbell" network
  - Improved efficiency (3.5 hrs vs 3 mins)

Utilization

- Serving last 12 months' load, requires:
  - 1064 nodes without time-sharing,
    - But only 168 nodes with time-sharing.
  - 19.1 years without space-sharing,
    - But only 1 year with space-sharing.

Outline

- Background and Related Work
- Experiment Life Cycle
- Efficiency and Utilization
- **New Experimental Techniques**
- Summary

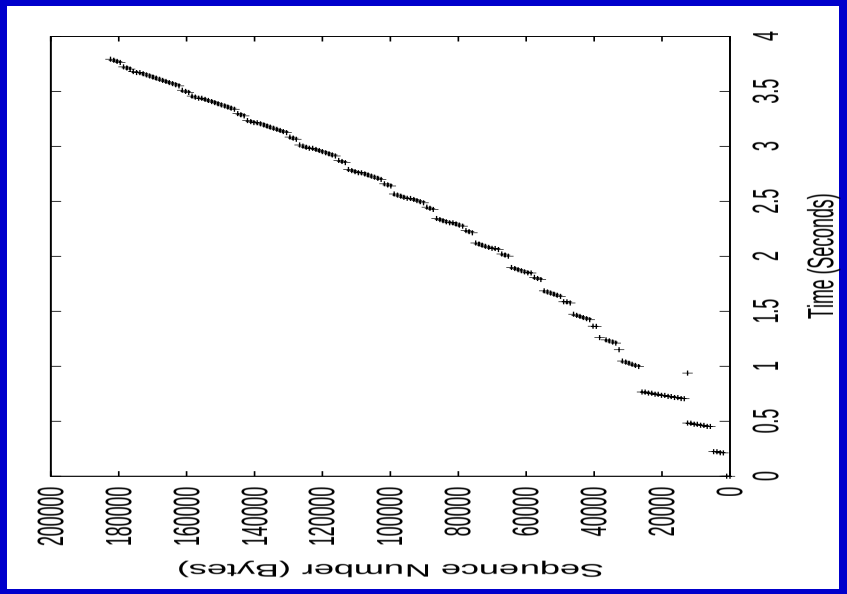
Parameter-Space Case Study

- Armada (Grid File System) Evaluation [Oldfield & Kotz'02]
- Run using batch experiments
- 7 bandwidths x 5 latencies x 3 application settings x 4 configs of 20 nodes
- 420 tests in 30 hrs (4.3 min apiece)

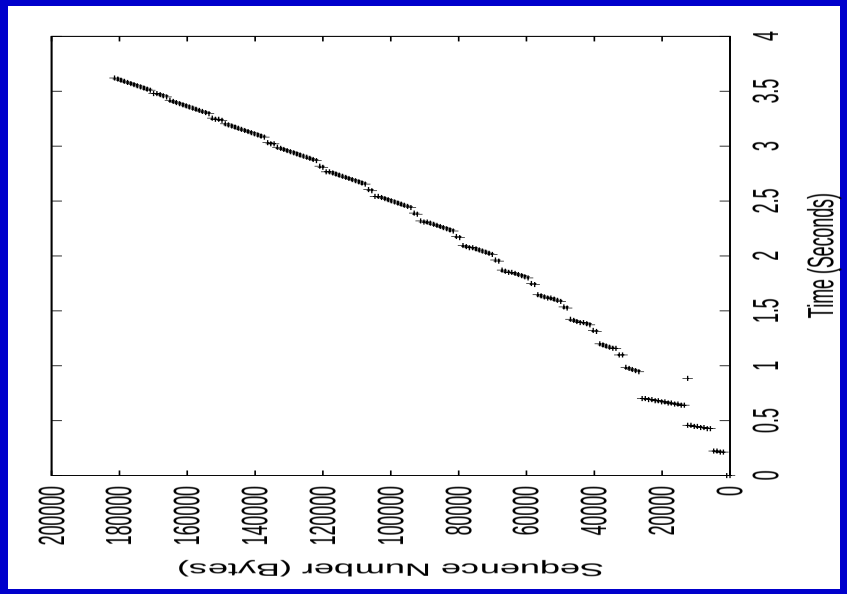
TCP Dynamics Case Study

- Runs ns regression tests on real kernels
- Compares empirical results vs. vetted simulation results
- Exploits simulation/emulation transparency to …
  - Check accuracy of simulation models, and …
  - Spot bugs in network stack implementations
- Infers packet loss from simulation output
- Injects failures into links via event system

TCP New Reno One Drop Test



ns



FreeBSD 4.5

Outline

- Background and Related Work
- Experiment Life Cycle
- Efficiency and Utilization
- New Experimental Techniques
- **Summary**

Beyond Experimentation …

- Today: Cluster management
  - Océano, Utility Data Centers, Cluster-on-Demand, …
- Future Work:
  - Reliability/Fault Tolerance
  - Distributed Debugging: Checkpoint/Rollback
  - Security "Petri Dish"

Summary

- Two orders of magnitude speedup in emulation setup and configuration time
- Provides a virtual machine for network experimentation
- Enables qualitatively new experimental techniques

www.netbed.org