Towards a Global-Scale Public Computer

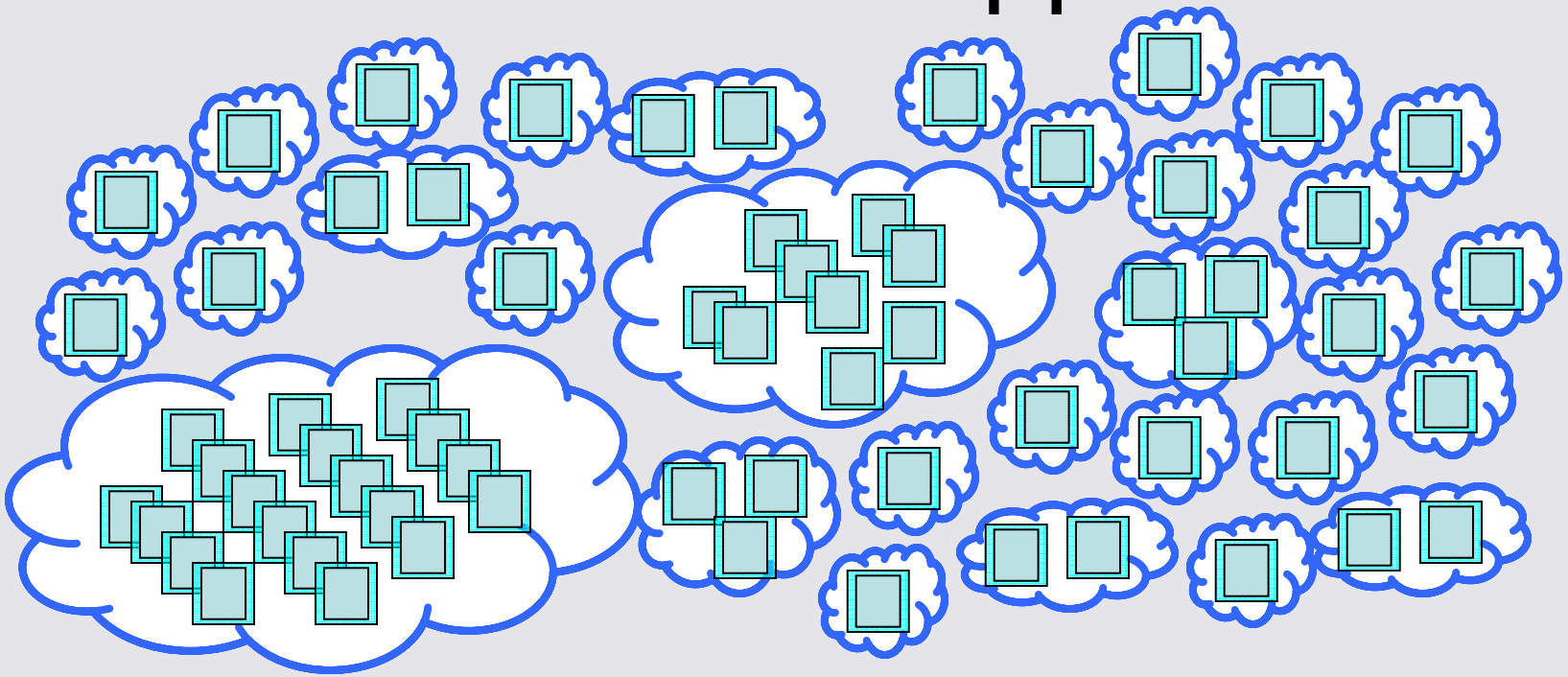
Project Kittyhawk at IBM Research

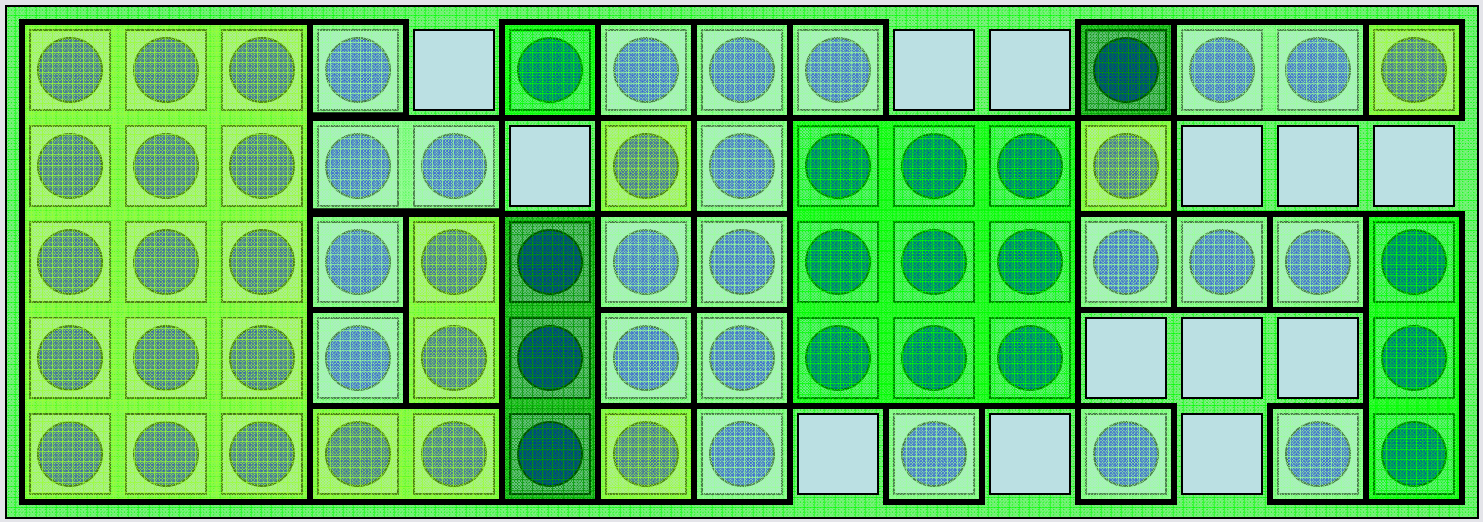
Presented by Jonathan Appavoo representing the Project Kittyhawk Team
Jonathan Appavoo, Volkmar Uhlig, Amos Waterland, Bryan Rosenburg
IBM T. J. Watson Research Center, New York

A widely accessible, scalable computer that represents a significant fraction of the world's current global computational capacity.

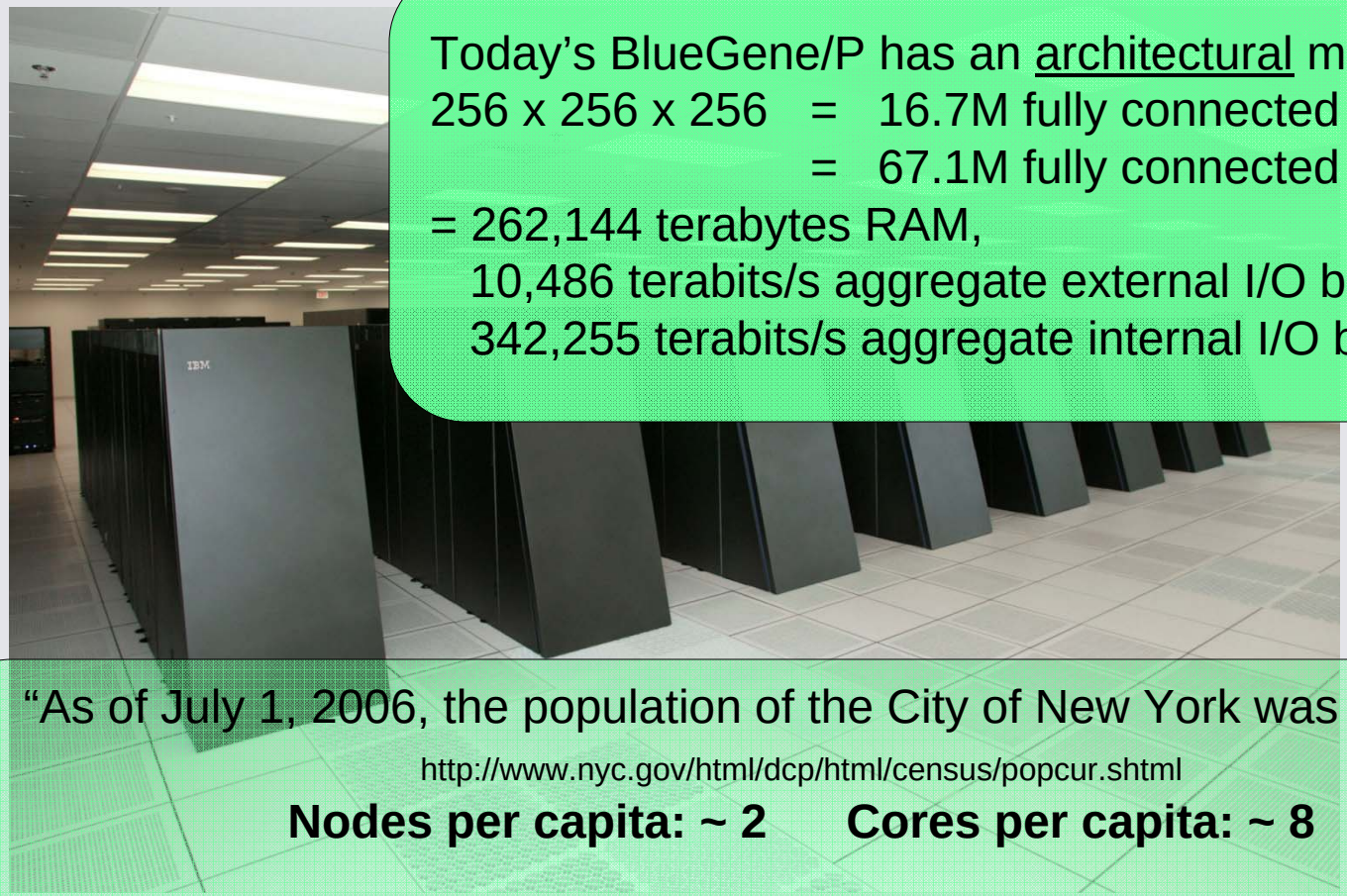
Kittyhawk
=
Global Computation
+
Massive Parallel Processor

Internet as an Application





Massive Parallel Processor



Today's BlueGene/P has an architectural maximum size of:
256 x 256 x 256 = 16.7M fully connected nodes
= 67.1M fully connected cores
= 262,144 terabytes RAM,
10,486 terabits/s aggregate external I/O bandwidth,
342,255 terabits/s aggregate internal I/O bandwidth.

"As of July 1, 2006, the population of the City of New York was 8,250,567"

http://www.nyc.gov/html/dcp/html/census/popcur.shtml

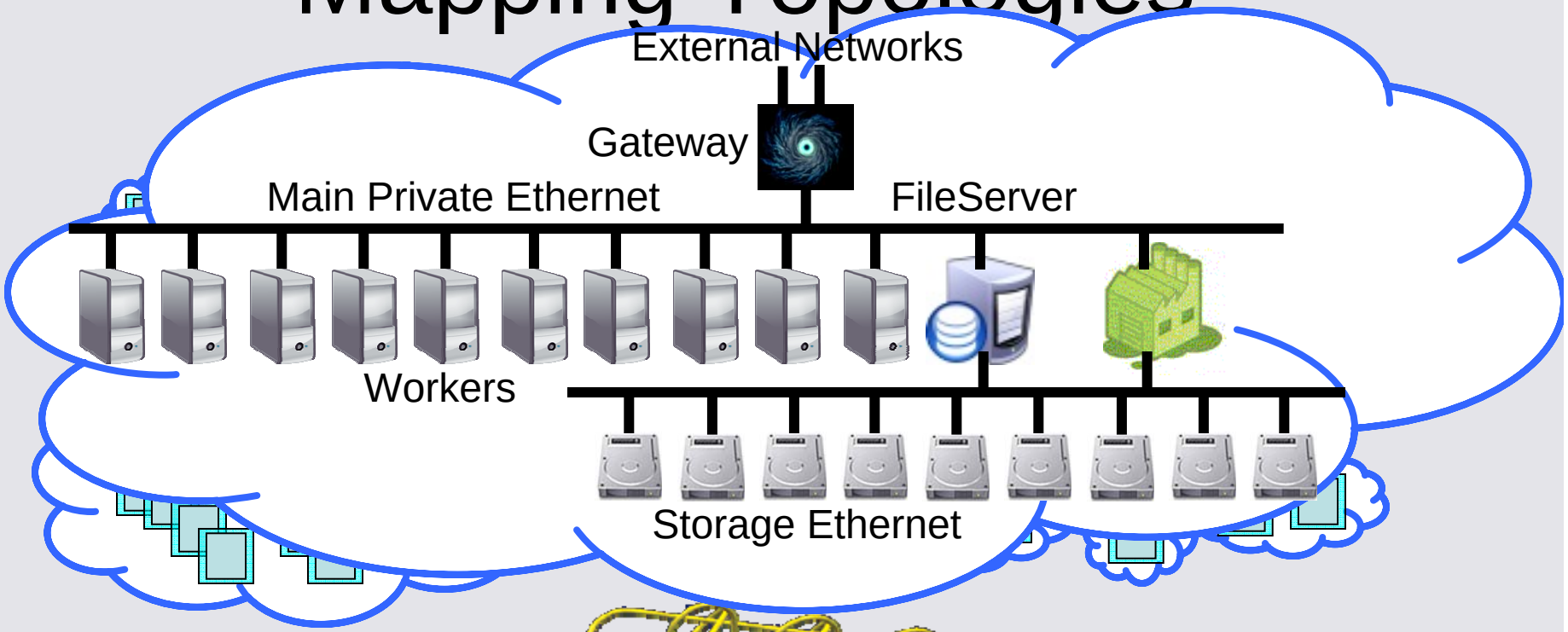
Nodes per capita: ~ 2 Cores per capita: ~ 8

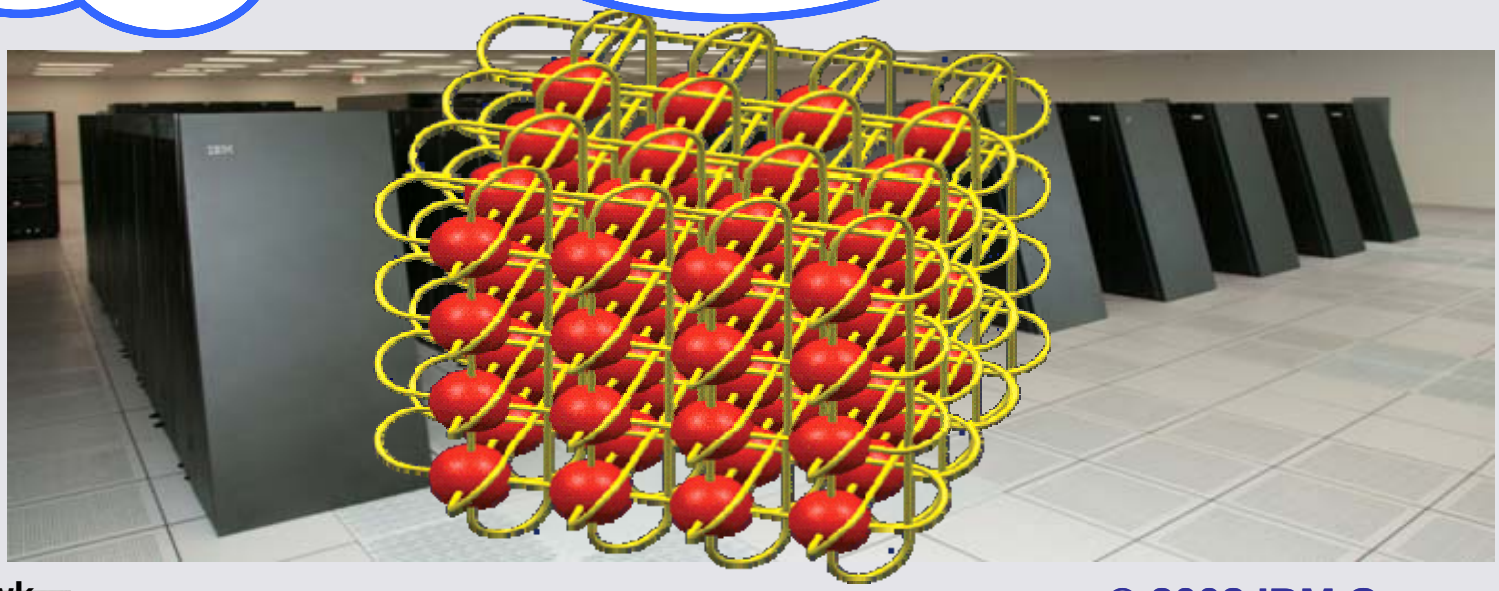
2005 Total US Volume Servers (<$25,000 per unit) = 9,897,000

Jonathan G. Koomey, "ESTIMATING TOTAL POWER CONSUMPTION BY SERVERS IN THE U.S. AND THE WORLD**",**Staff Scientist, Lawrence Berkeley National Laboratory and Consulting professor, Stanford University,Final report February 15, 2007

Nodes per server: ~ 1.6

Mapping Topologies



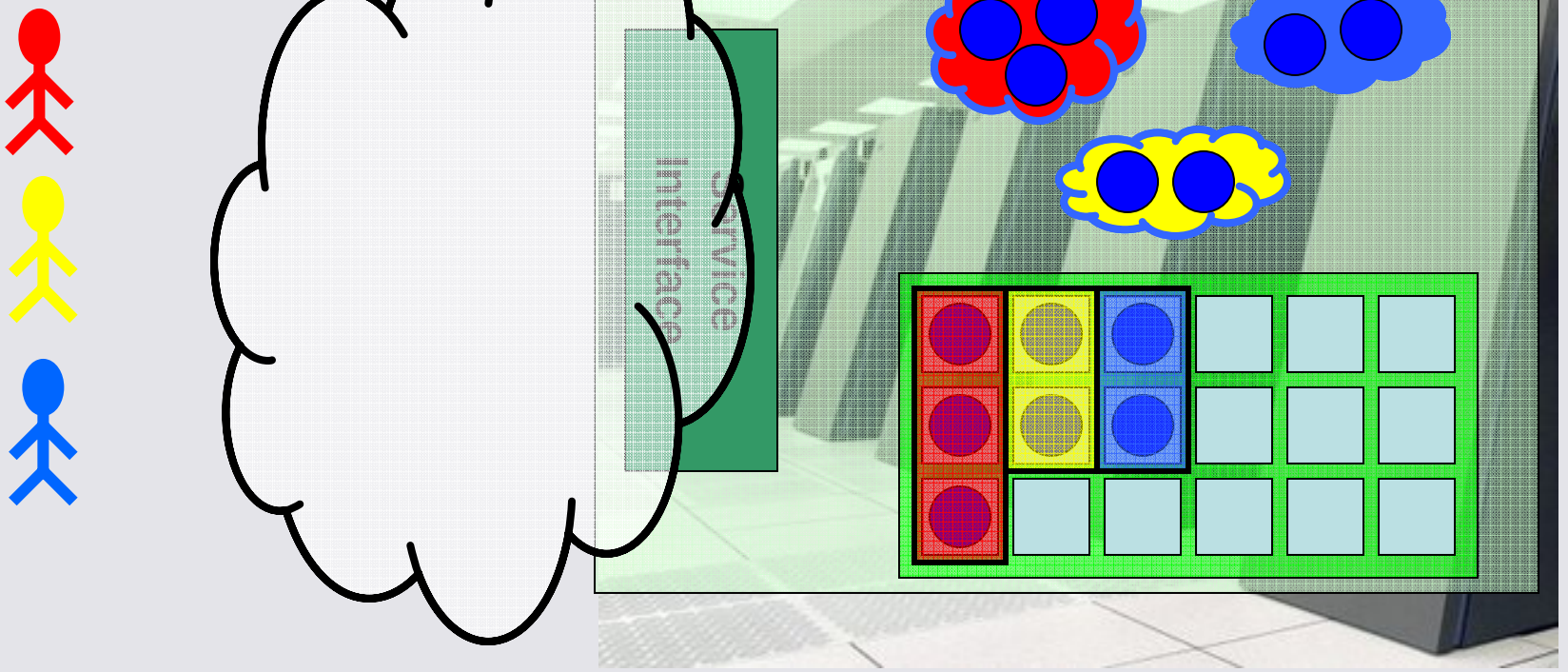


Abstract System View

Users (developers, service providers, resellers)

Public Access Network

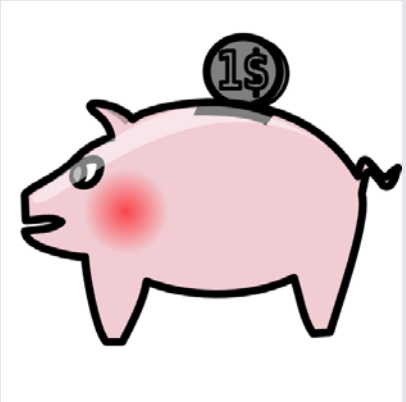
The Computer



The Key Points of this Talk

1. Large-scale communication-centric system through **Aggressive Integration.**
2. Fairness through **Raw Hardware Access.**
3. Competition and cooperation through **Dynamic, Hardware-enforced Communication Domains**

Benefits of the Approach



Outline

1. A global computer & global computation.
2. Our prototype.
3. Why hardware-centric? When virtual is bad!
4. Our digital future -- by accident or design?

Our Take on a Global-scale Computer and Global Computation

A Global Scale Computer: A well-specified public, “software-less”, massively-parallel system, on which users can construct services, of arbitrary scale, out of metered and billed common units of its capacity grouped in domains of communication they specify and control.

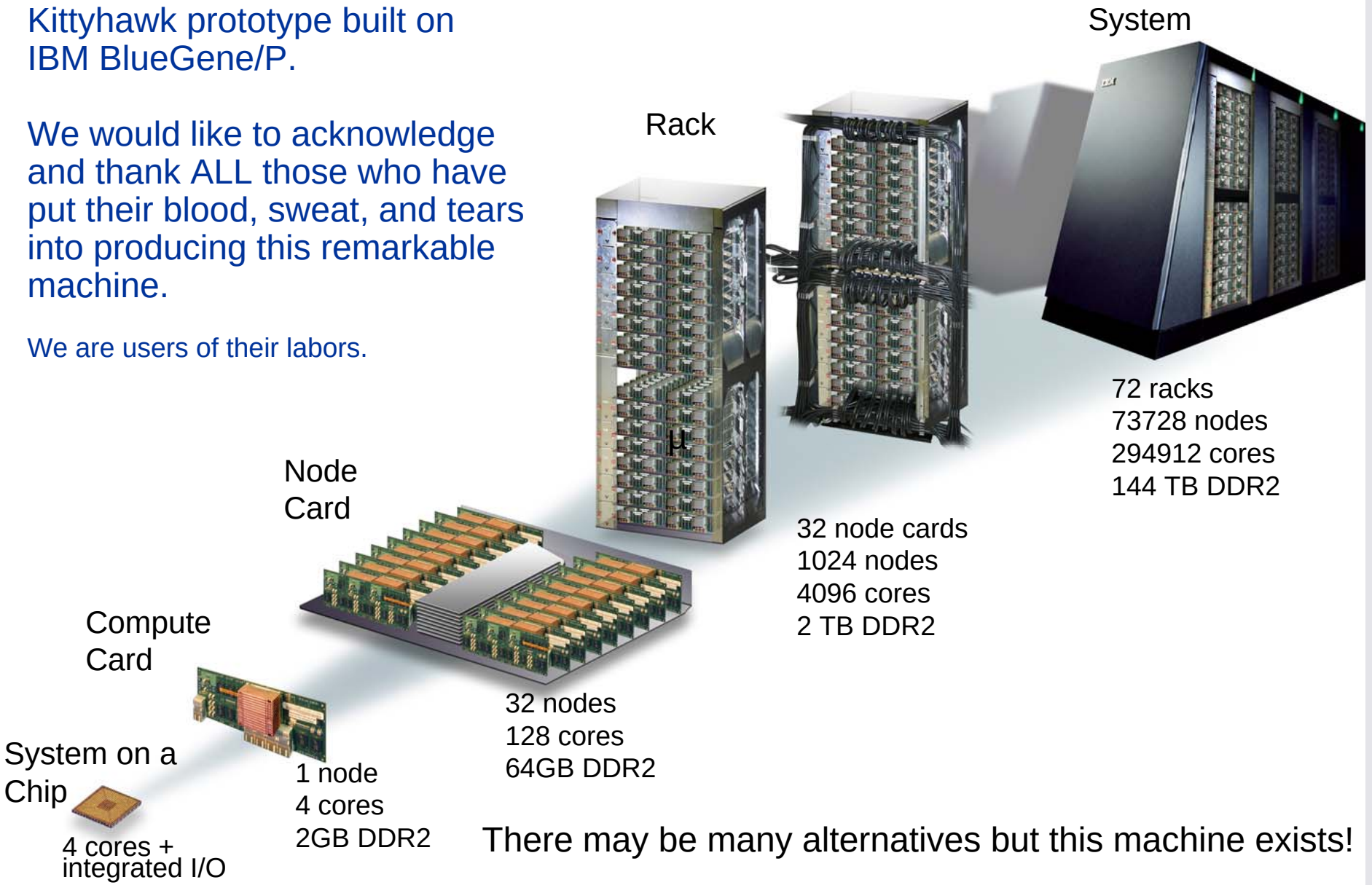
Our Prototype

1. Massive parallel processor: BlueGene/P
2. Evolution toward a global-scale computer
3. Demos

Kittyhawk prototype built on IBM BlueGene/P.

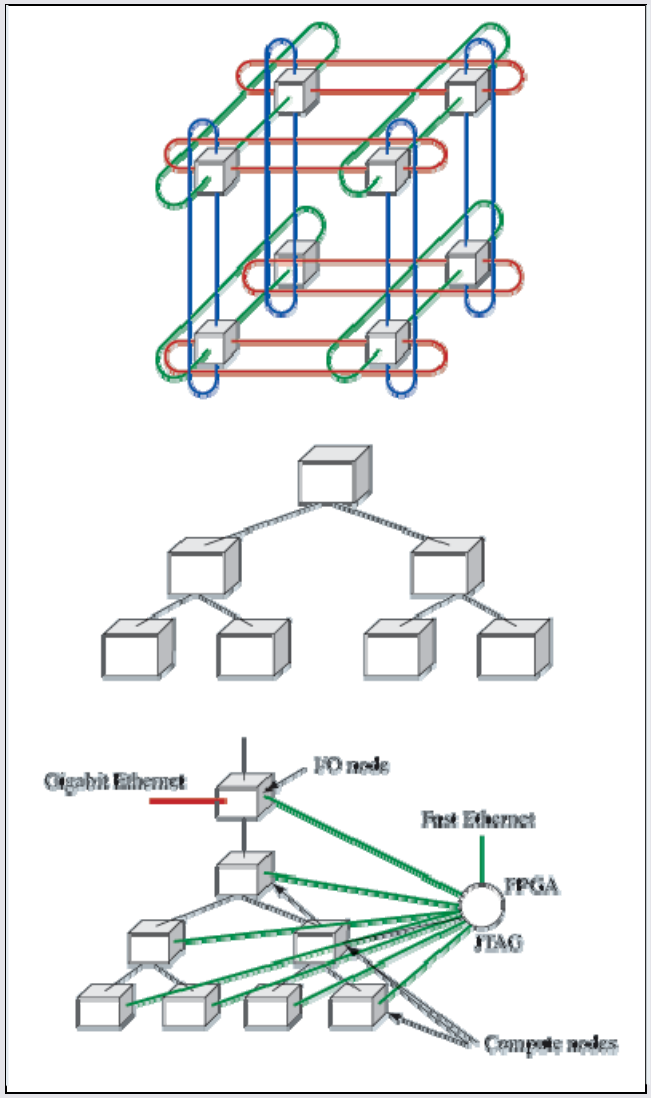
We would like to acknowledge and thank ALL those who have put their blood, sweat, and tears into producing this remarkable machine.

We are users of their labors.



There may be many alternatives but this machine exists!

5 for the price of 1



1) 3 Dimensional Torus

- Interconnects all compute nodes.
- Adaptive cut-through hardware routing.
- 3.4 gigabits/s on all 12 node links (5.1 gigabits/s per node).
- 0.5 µs latency between nearest neighbors.
- 1.7/2.6 terabits/s bisectional bandwidth, 188 terabits/s total bandwidth (72k machine).

2) Collective Network

- One-to-all broadcast functionality.
- Reduction operations for integers and doubles.
- 6.8 gigabits/s of bandwidth per link per direction.
- Latency of one-way tree traversal 1.3 µs.
- Interconnects all compute nodes and I/O nodes.
- ~62 terabits/s total binary tree bandwidth (72k machine).

3) Low Latency Global Barrier and Interrupt

- Latency of one-way signal to all 72K nodes 0.65 µs.

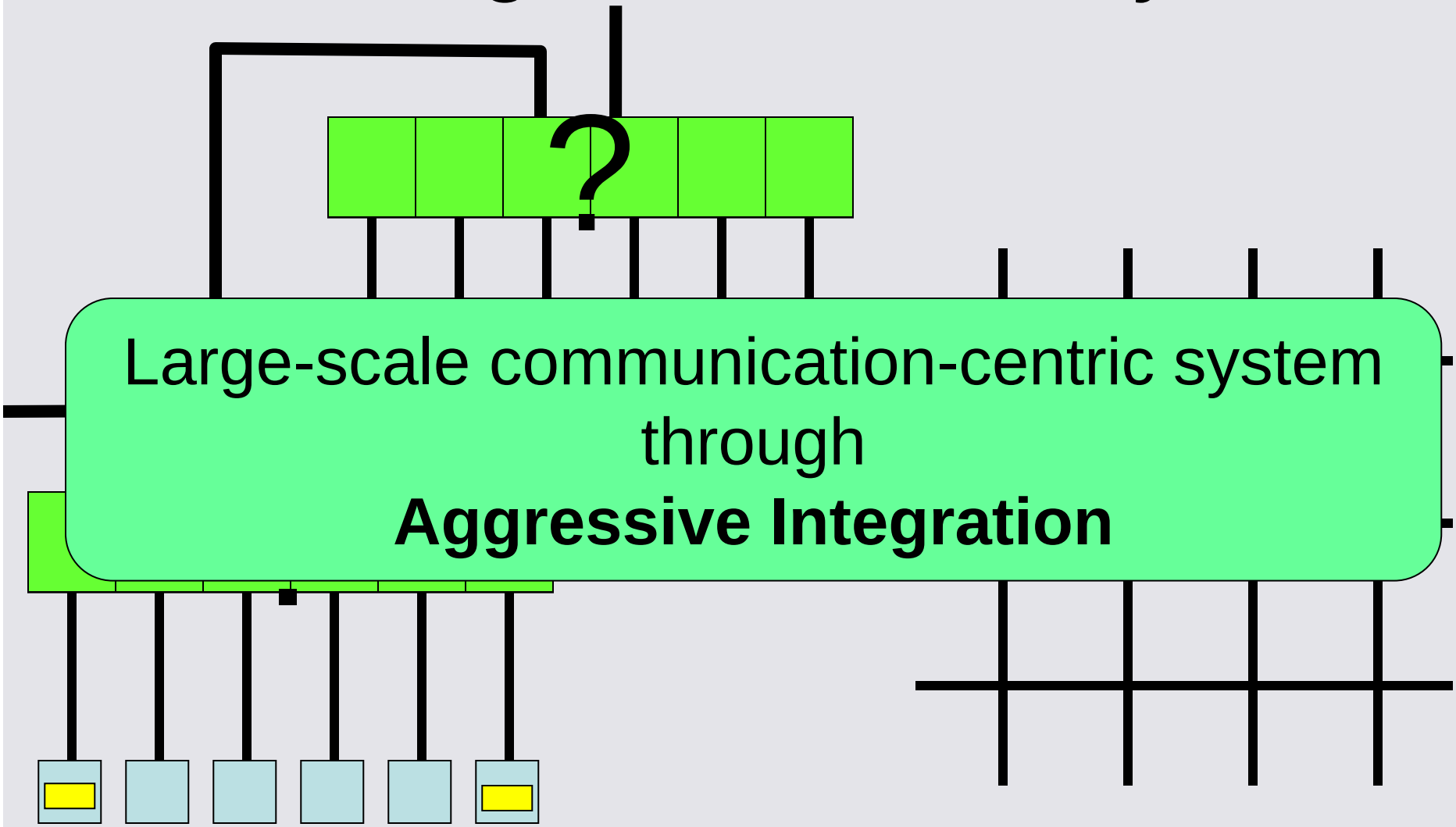
4) 10Gig Functional External I/O Ethernet Network

- I/O nodes only.

5) 1Gig Private Control Ethernet

- Provides JTAG access to hardware.
- Accessible only from Secure Service Node System.

Integrated Scalability



Massively Scalable Networking

Every node in the system can be identified by a unique x,y,z coordinate:

1. Fabric scales with nodes: nodes ***form*** the fabric.
2. Cut-through routing: fabric cooperates to route a message point-to-point with ***no*** software involvement.
3. Global, symmetric, and transparent: spans all hardware domains; node cards, mid-planes, and ***racks***.

Scale

Racks	1	8	64	256
Nodes	1,024	8,192	65,536	262,144
Cores	4,096	32,768	262,144	1,048,576
Internal aggregate bandwidth gigabits/s	20,890	167,117	1,336,934	5,347,738
External aggregate bandwidth gigabits/s	640	5,120	40,960	163,840
RAM in terabytes (current limit / architectural max)	2 **(4/16)**	16 **(32/128)**	128 **(256/1024)**	512 **(1024/4096)**
Peak power consumption (kW)	35	280	2240	8960

"Available in configurations ranging from one to 256 racks, Blue Gene/P is the innovative new solution from IBM to further expand the limits of breakthrough science without sacrificing efficiency."

http://www.ibm.com/systems/deepcomputing/bluegene/

Space and Power Consumption for a Million Cores

Mega Data Center with Standard Servers

- Power: 50 MW (~ $500 million)
- Floor space: 500,000 square feet

BlueGene/P

- Power: 10 MW (~ $100 million)
- Floor space: 10,000 square feet

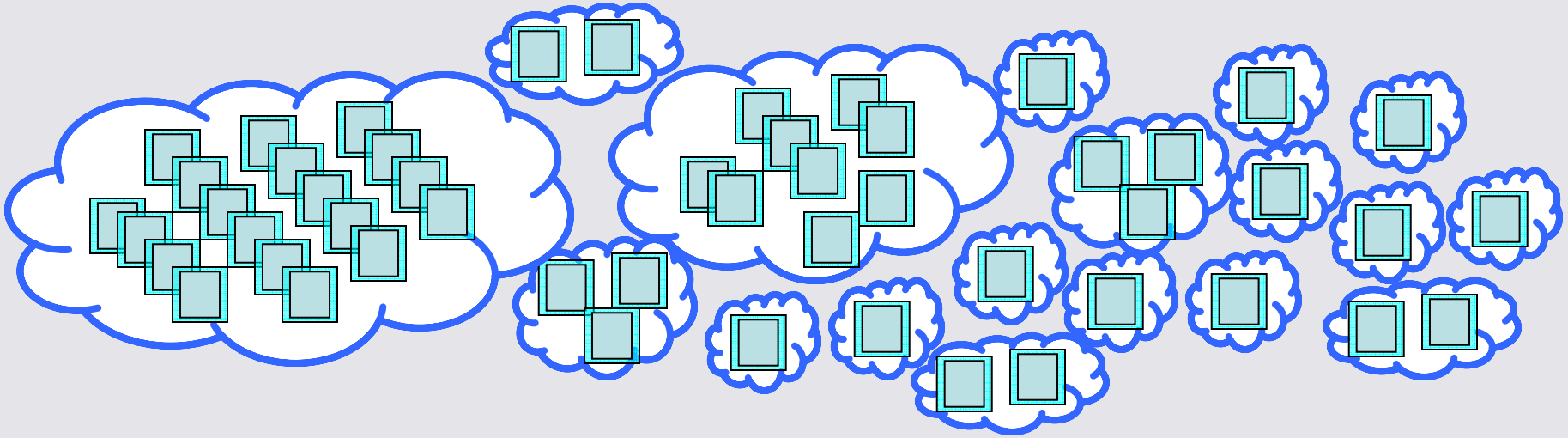
Evolution towards a Global-scale Computer

"A parallel computer with a billion processors might provide a computational utility analogous to existing gas and electric utilities."

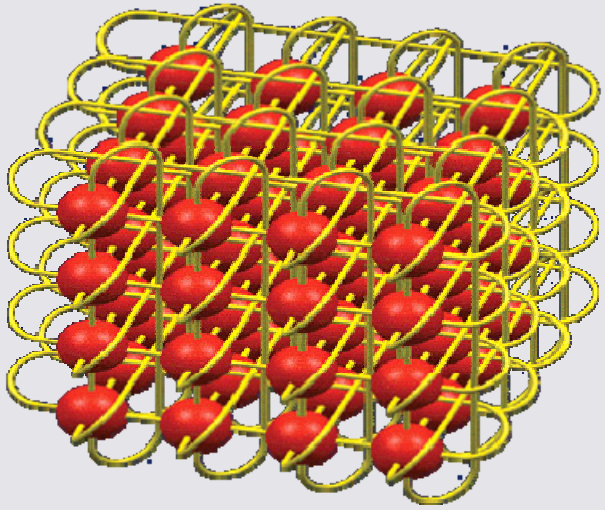
W. Daniel Hillis, "The Connection Machine",

Scientific American 256, 6 (June 1987),108--115.

Supporting Global Computation



Large-scale Communication-centric System



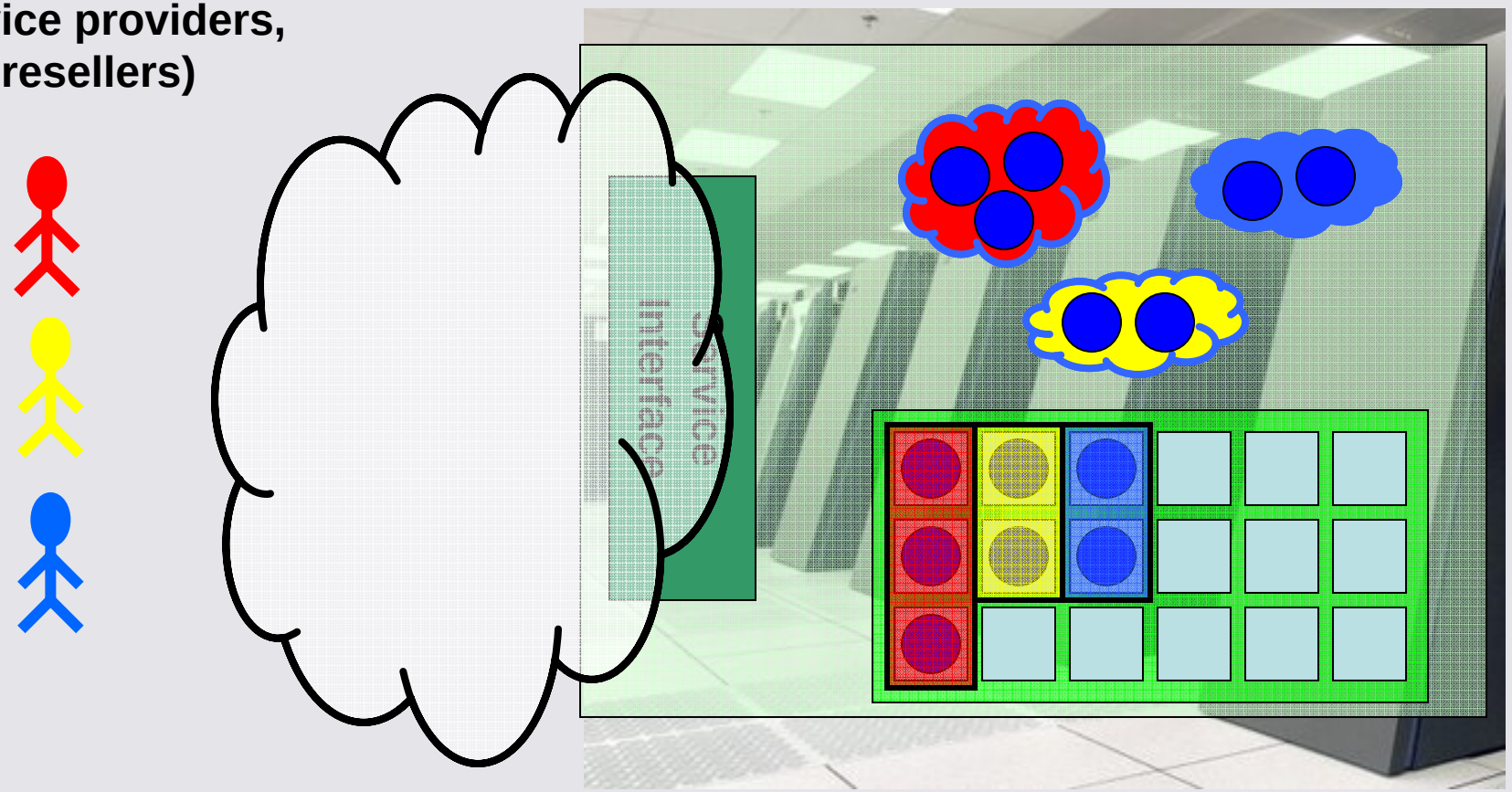
Aggressive Integration

Abstract View

Users (developers, service providers, resellers)

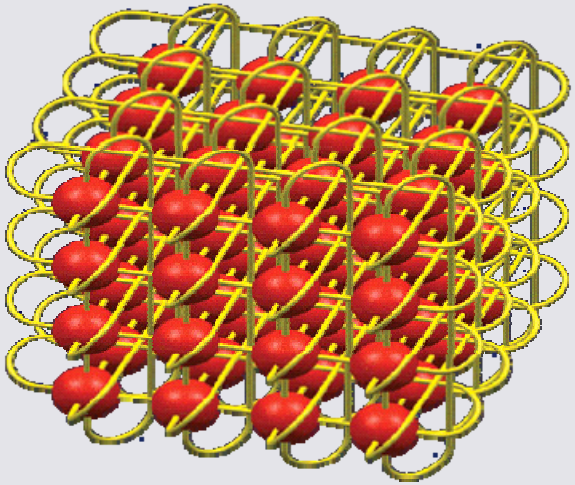
Public Access Network

The Computer



A Global-scale Computer

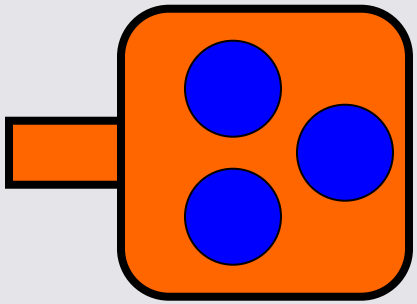
Large-scale Communication-centric System



Aggressive Integration

+

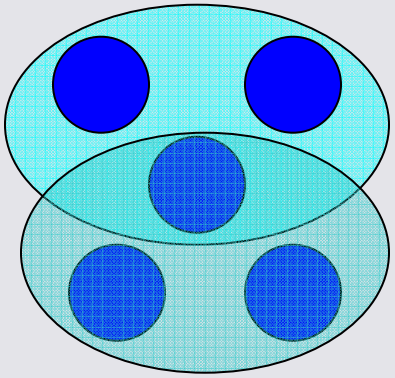
Control Channels



Raw Hardware Access

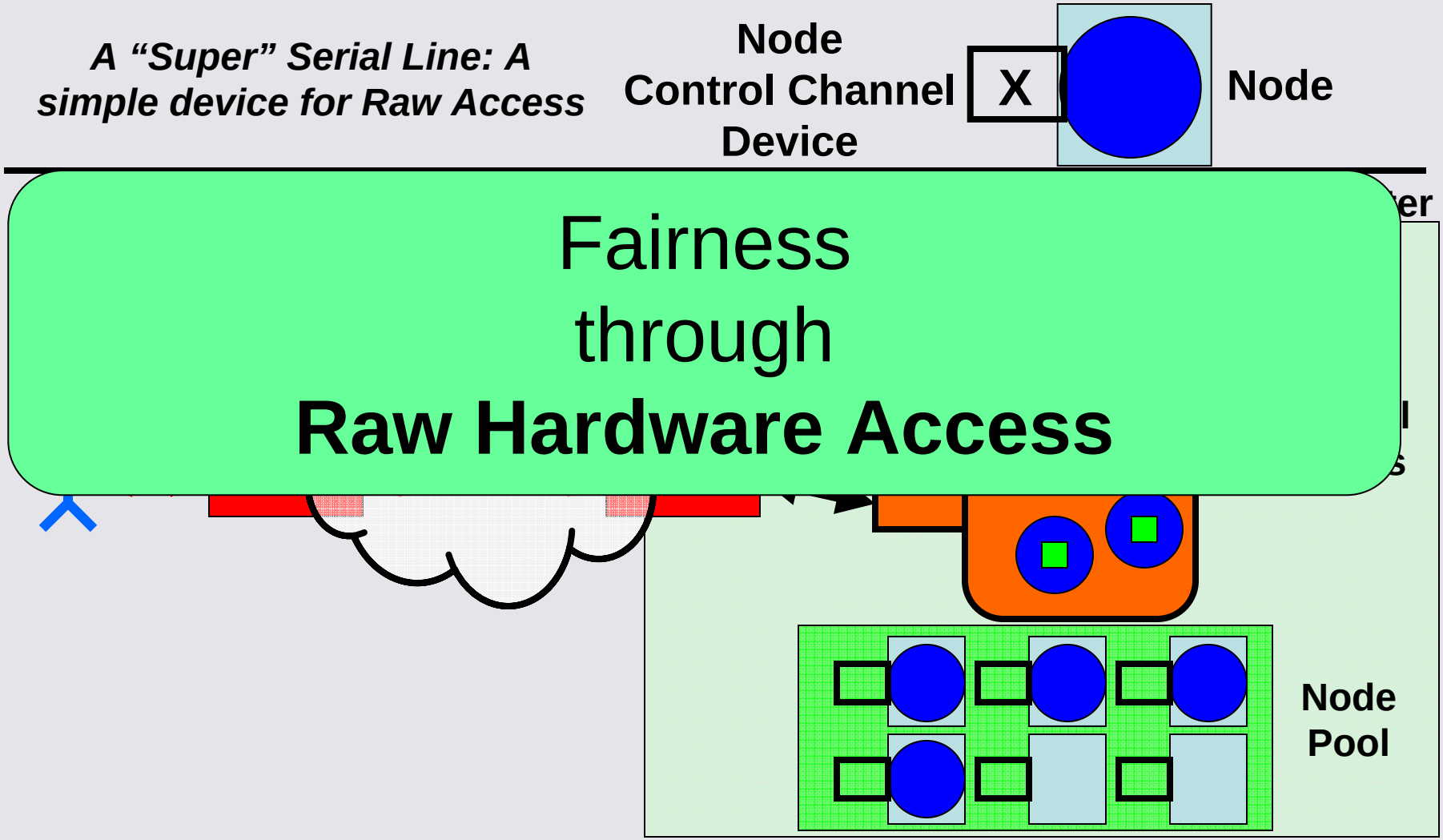
+

Communication Domains



Primitive for Competition & Cooperation

Public Access Control Channel as a Multicast Console Device

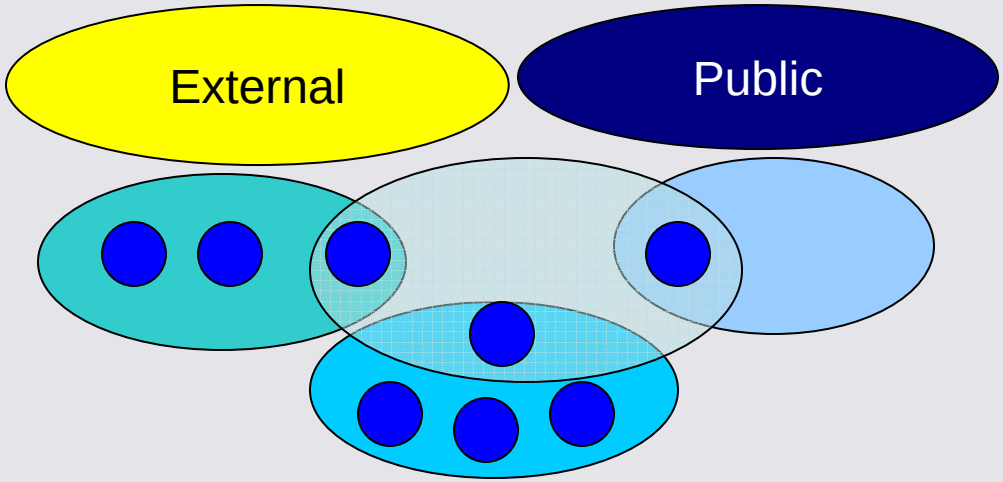


Communication Domains as a Core System Primitive

Domains: Hardware-enforced groups of HW endpoints that can communicate.

1. Power in realizing socio-economic relationships in topologies.
2. Individuals and organizations have freedom to isolate and protect resources and thus their interests and property.
3. Individuals and organizations can choose to cooperate by “simply” instantiating a resource in the public communication domain or a shared domain.

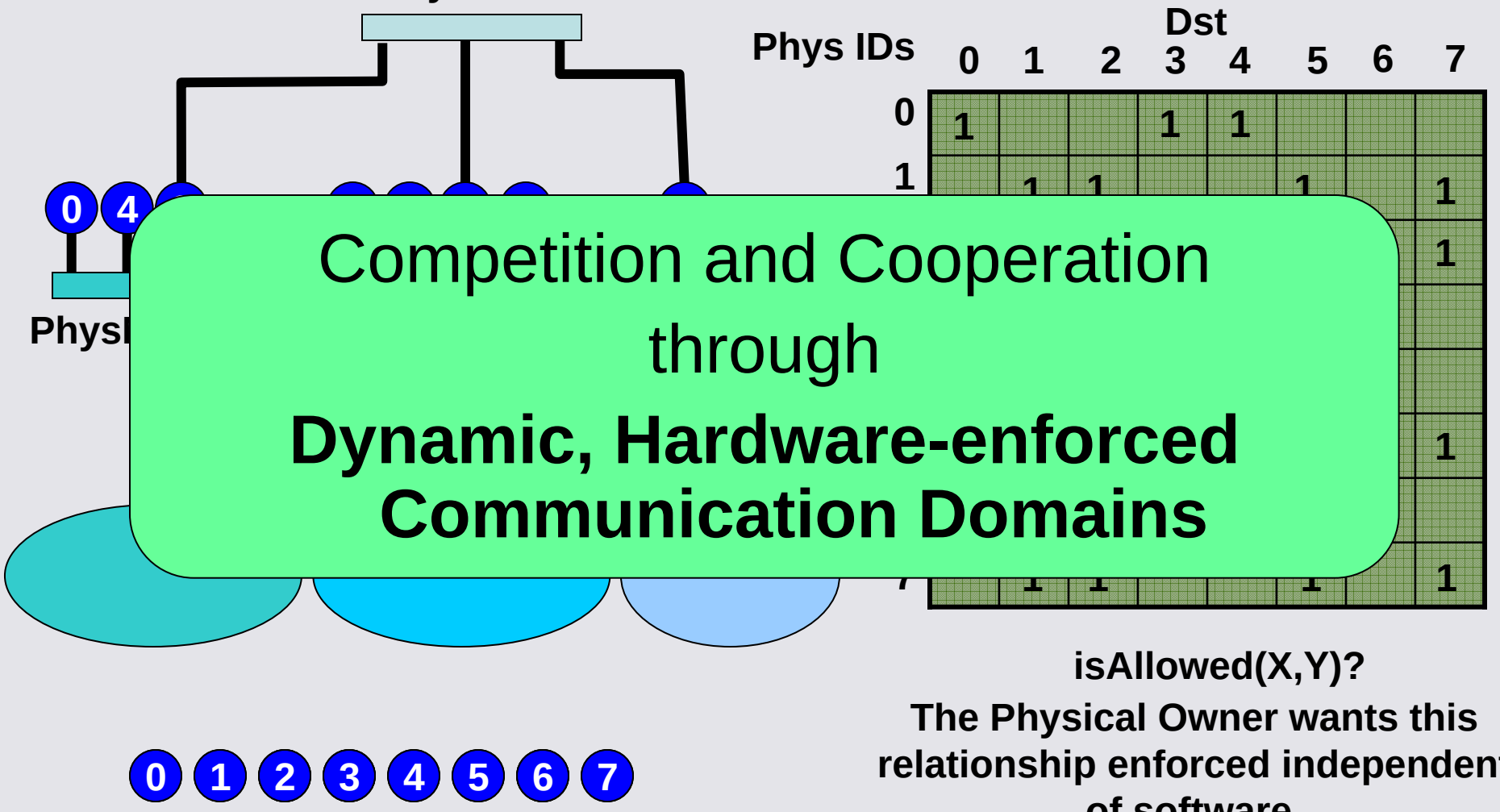
Domains + HW Endpoints



Hardware enforces communication constraints; software controls everything else.

Software defines how and what is communicated: raw bus use, cache lines, pages, ethernet frames, etc.

Communication Domains as Permission



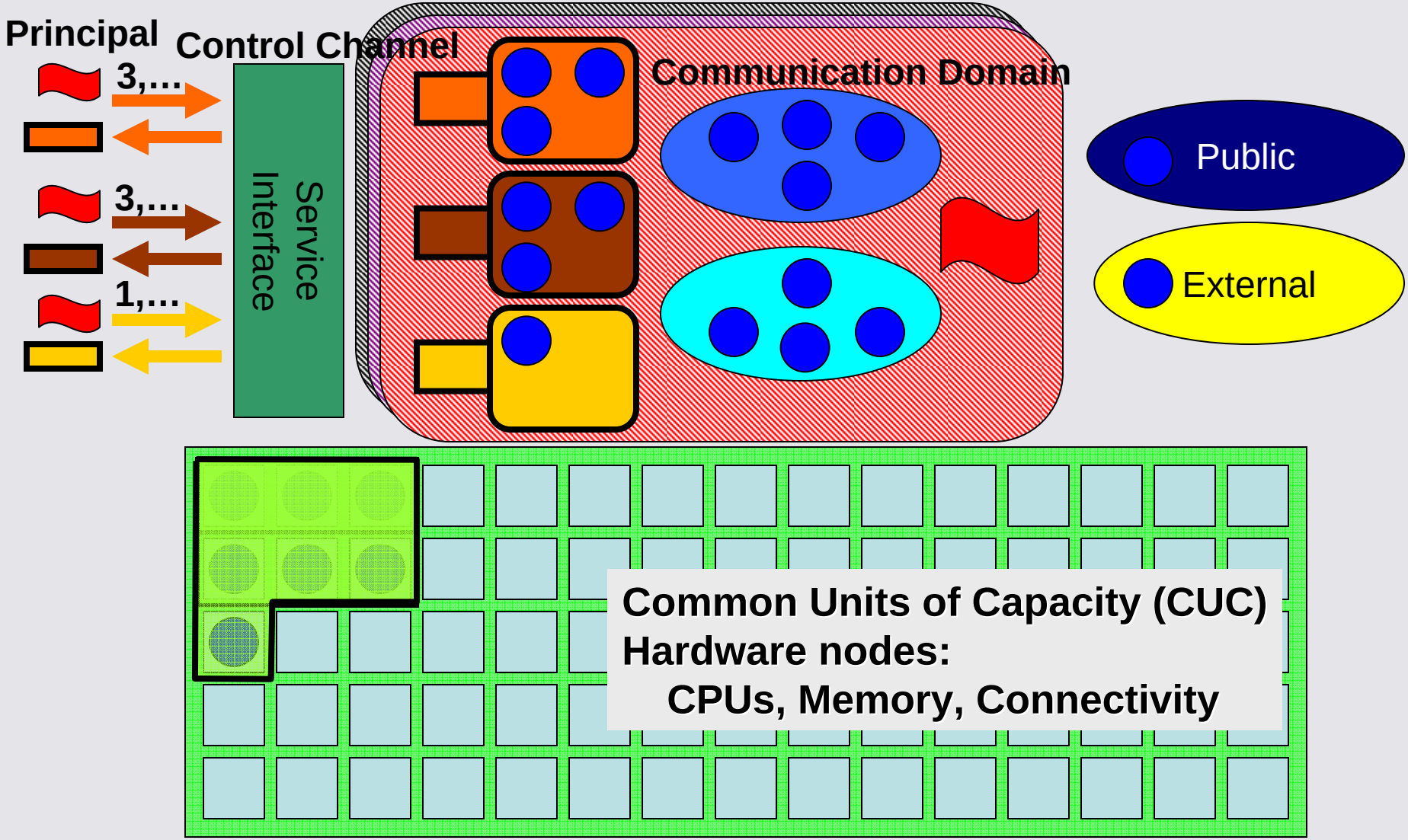
Competition and Cooperation through

Dynamic, Hardware-enforced Communication Domains

isAllowed(X,Y)?

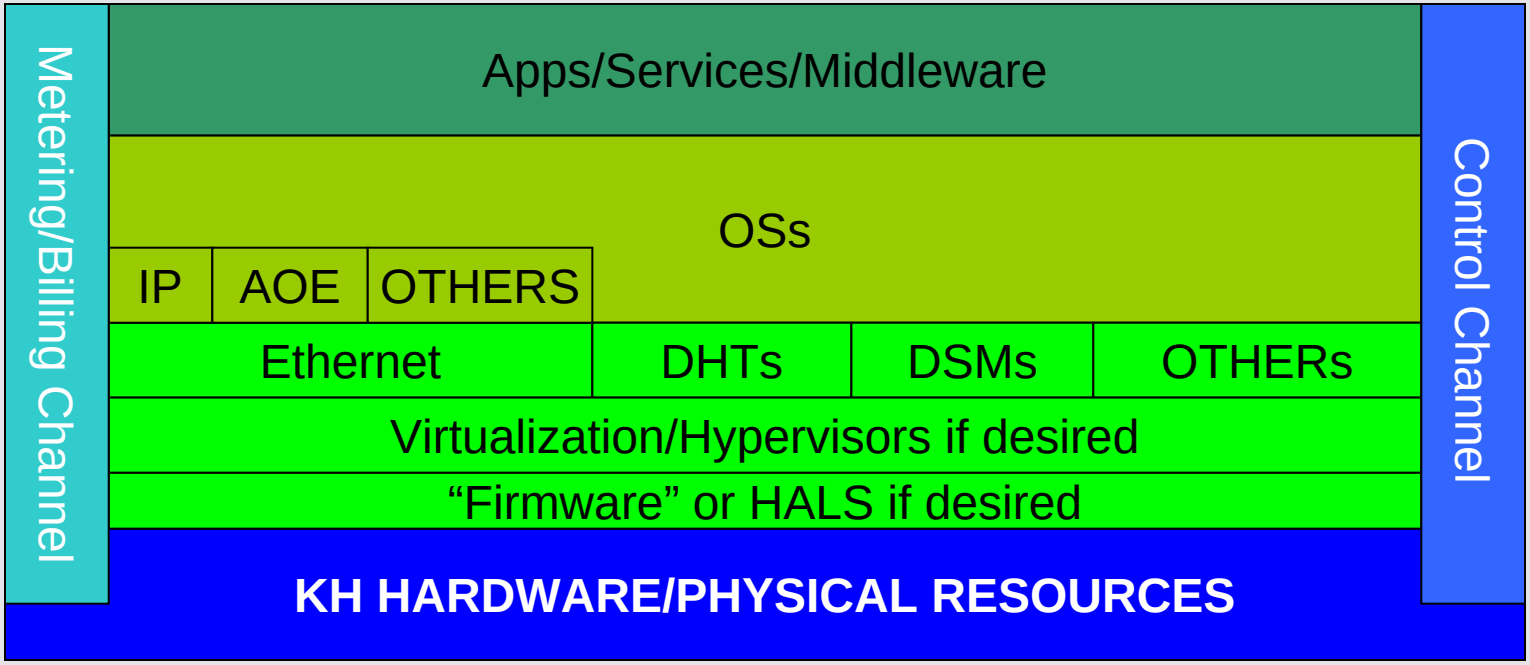
The Physical Owner wants this relationship enforced independent of software.

CUC, Principals, Control Channels, Communication Domains



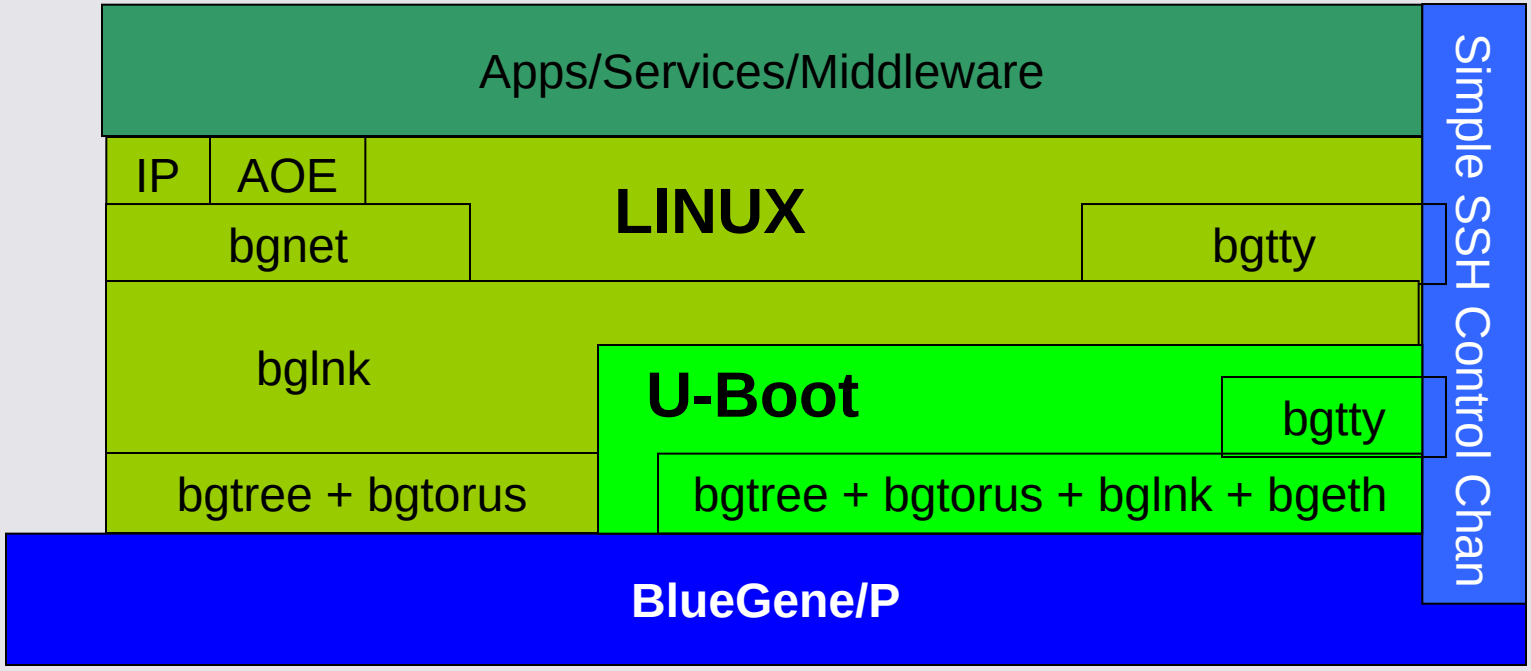
An Online Stack of Choice and Composition (SW as Refinement)

Anyone can revolutionize usage with new software at any level.
Anyone can create new environments.
Anyone can address niche needs via niche proprietary solutions.



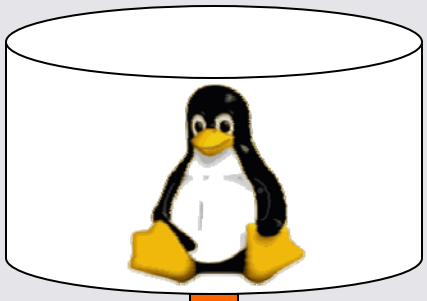
Exploring Kittyhawk via a Prototype on BlueGene/P

Current exemplar prototype built on BG/P hardware and predominately open-source software.



Trying out L4 (http://www.l4ka.org/) for both hardware prototyping and example virtualization.

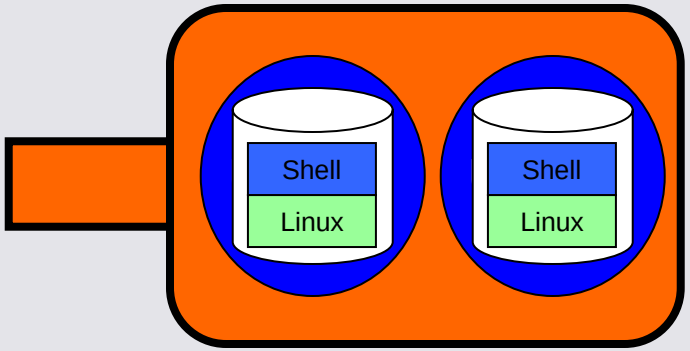
Bootstrapping: Open-source Software Appliances

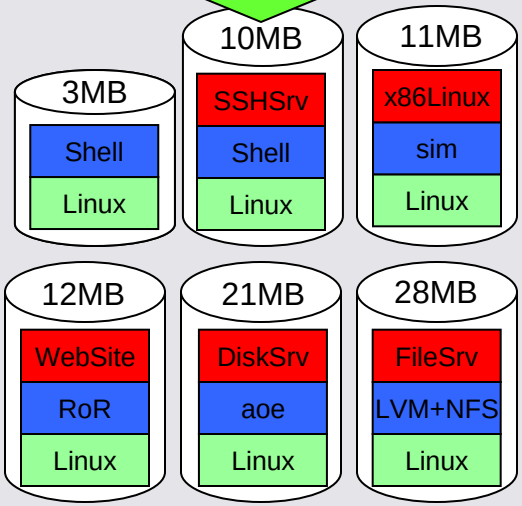


Full Linux Install Image (Distribution)

Extract

Trace & Package





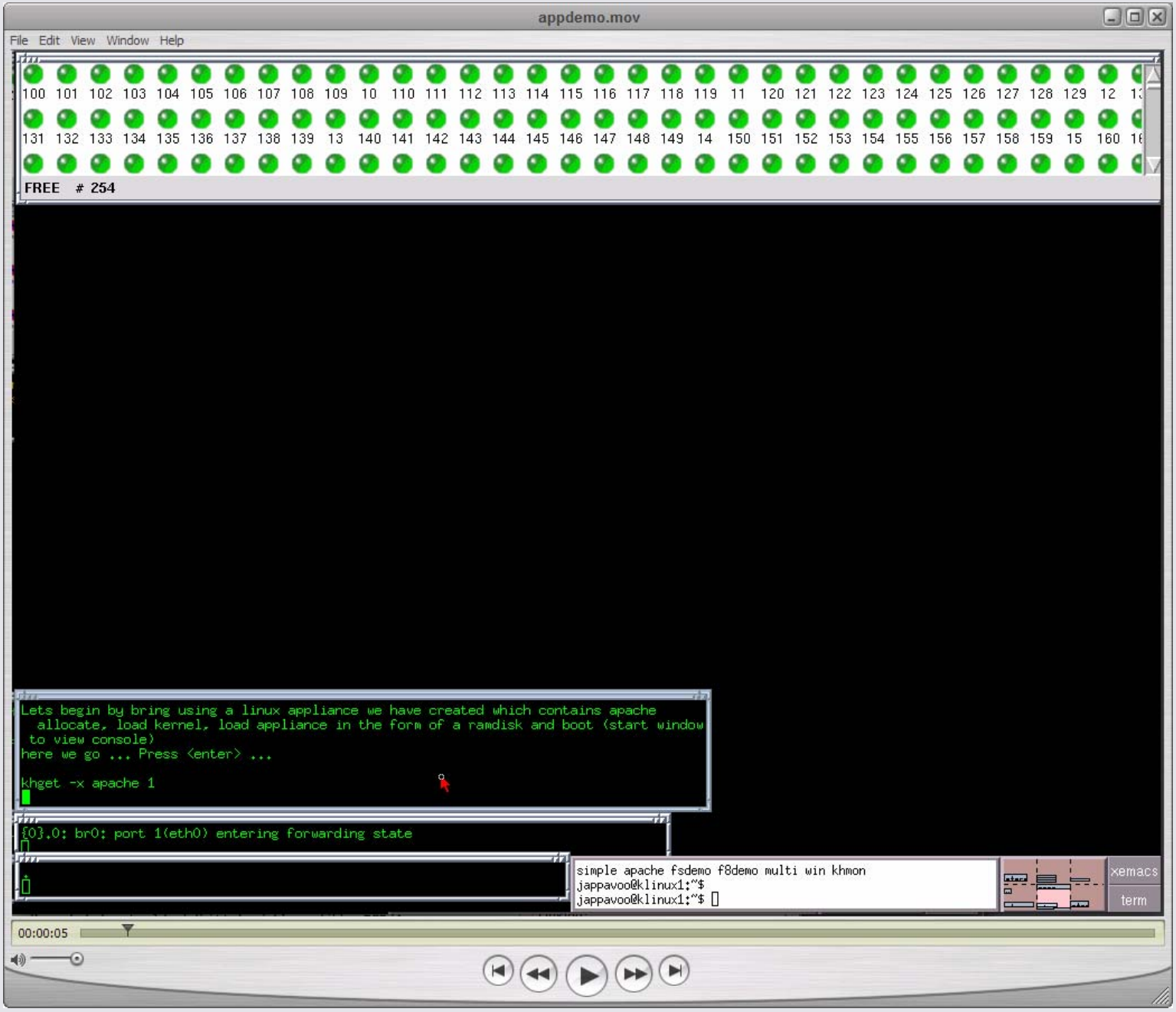
Bootable Ramdisks

Demos

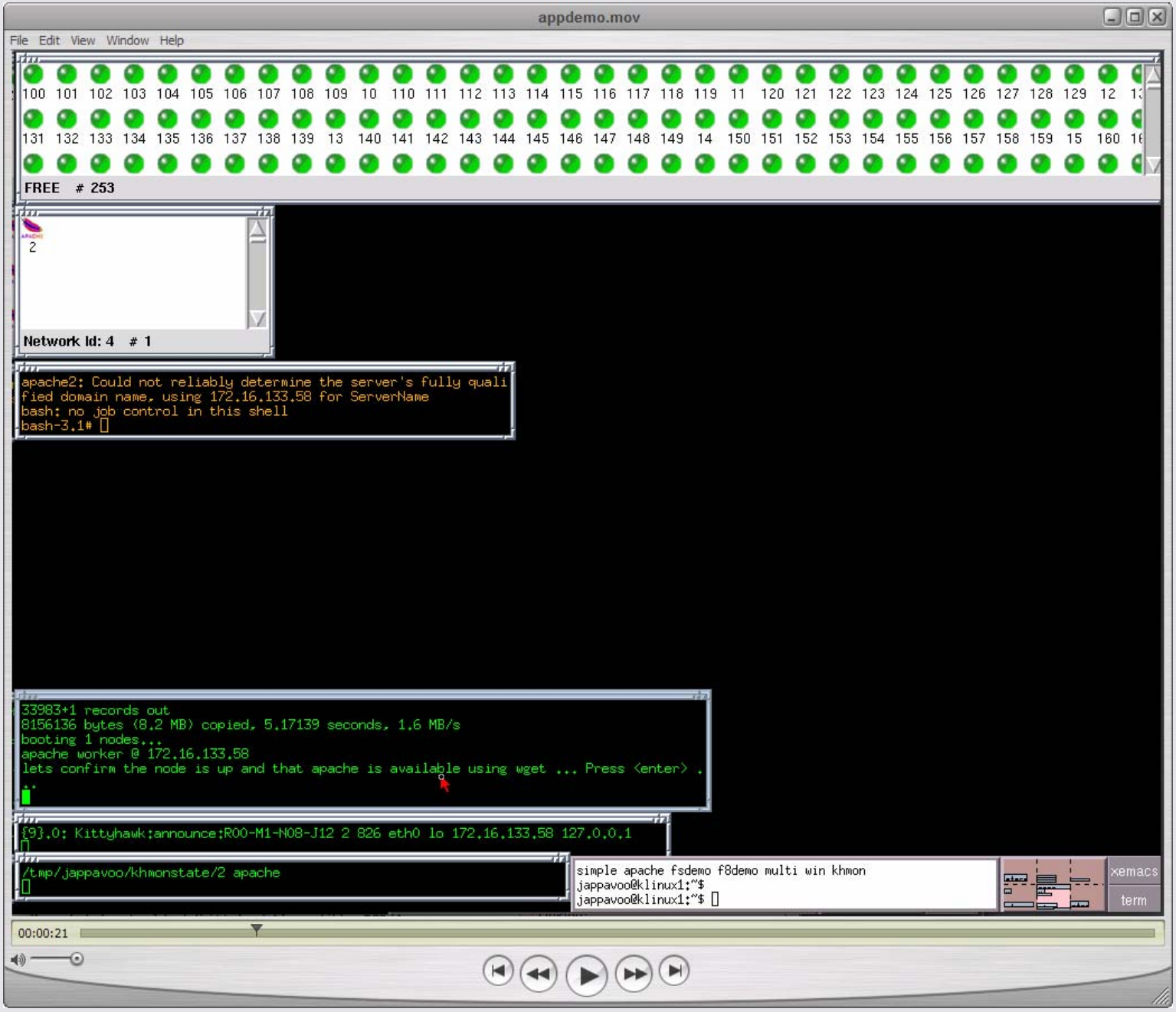
Building a Web Server Farm

http://www.research.ibm.com/kittyhawk/movies/appdemo.mov

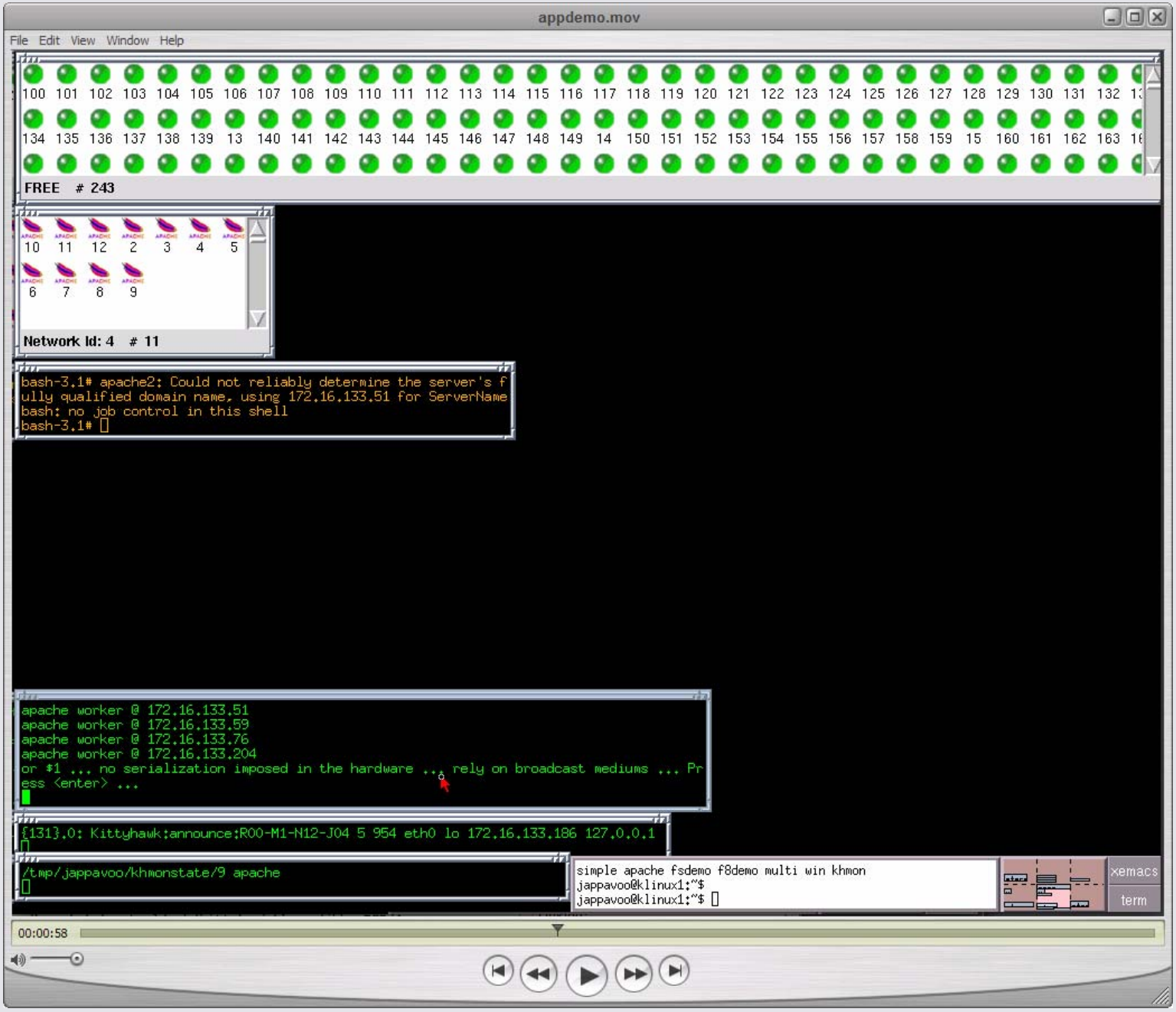
00:05



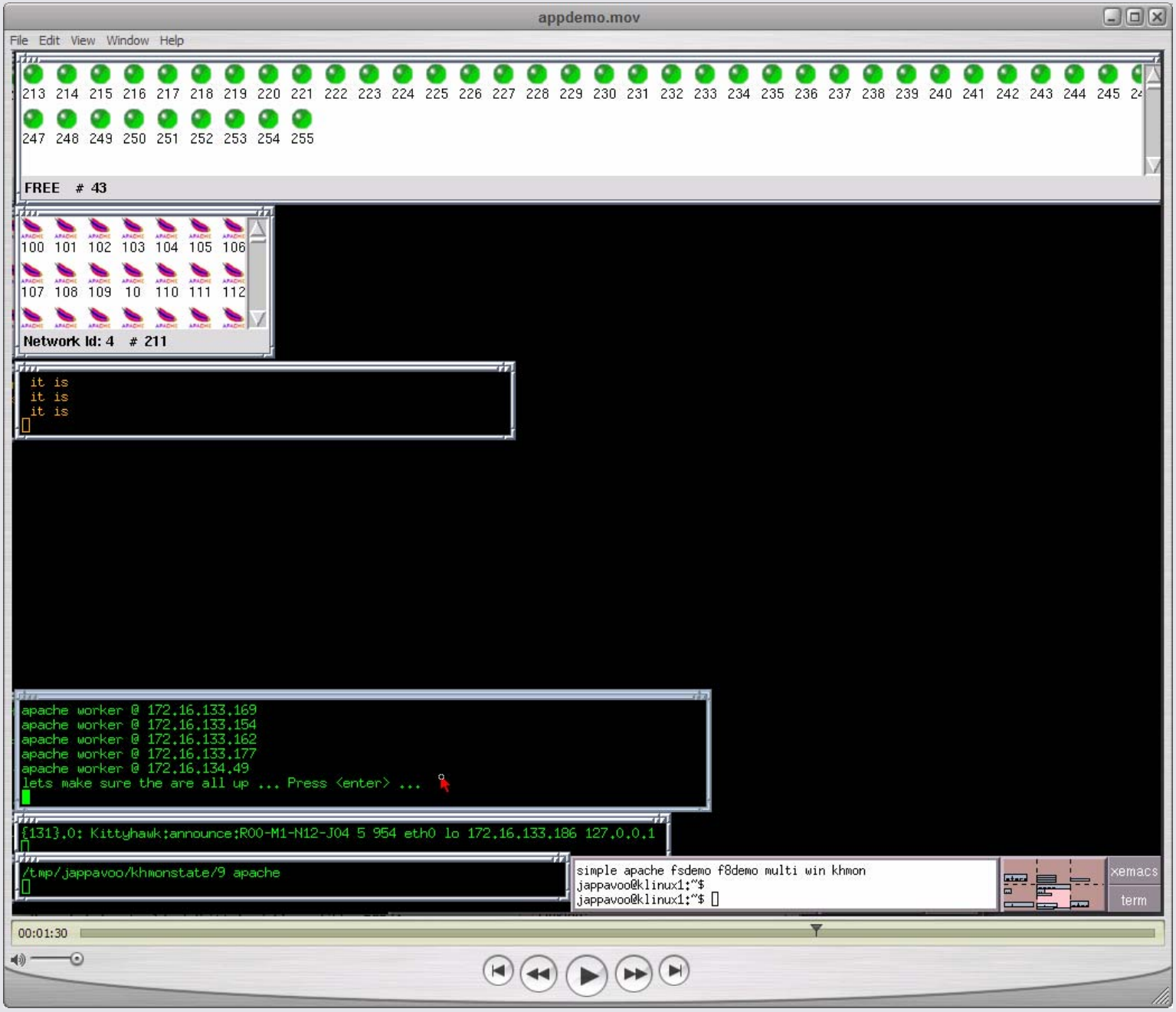
00:21



00:58



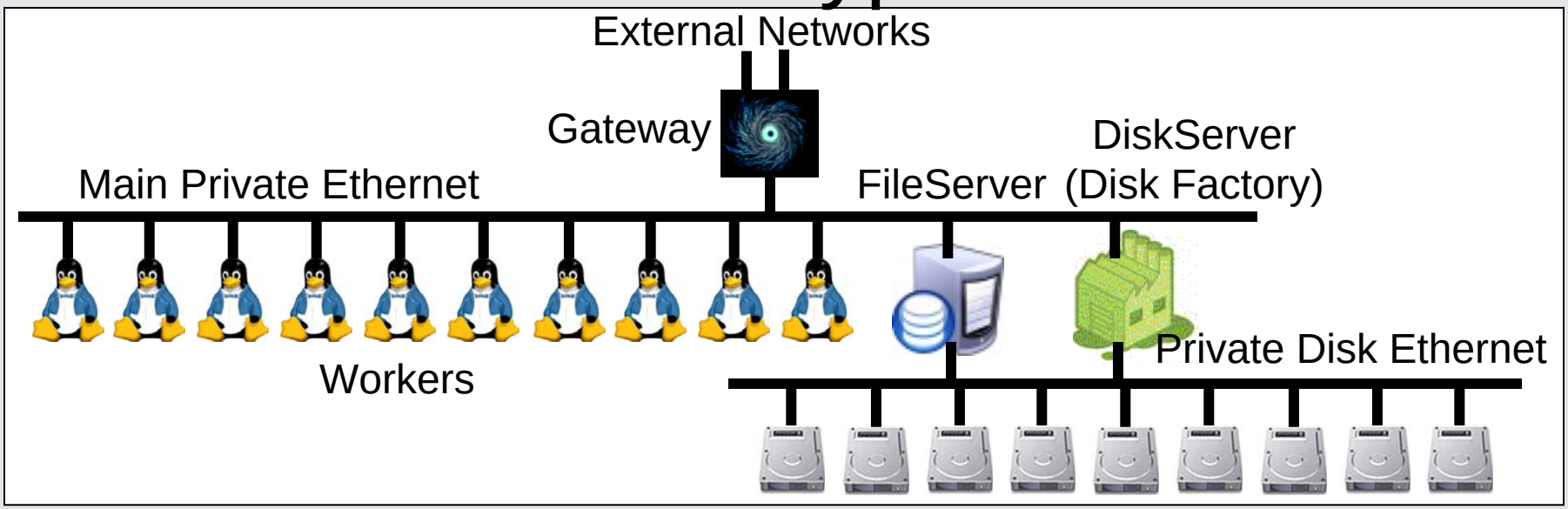
01:30

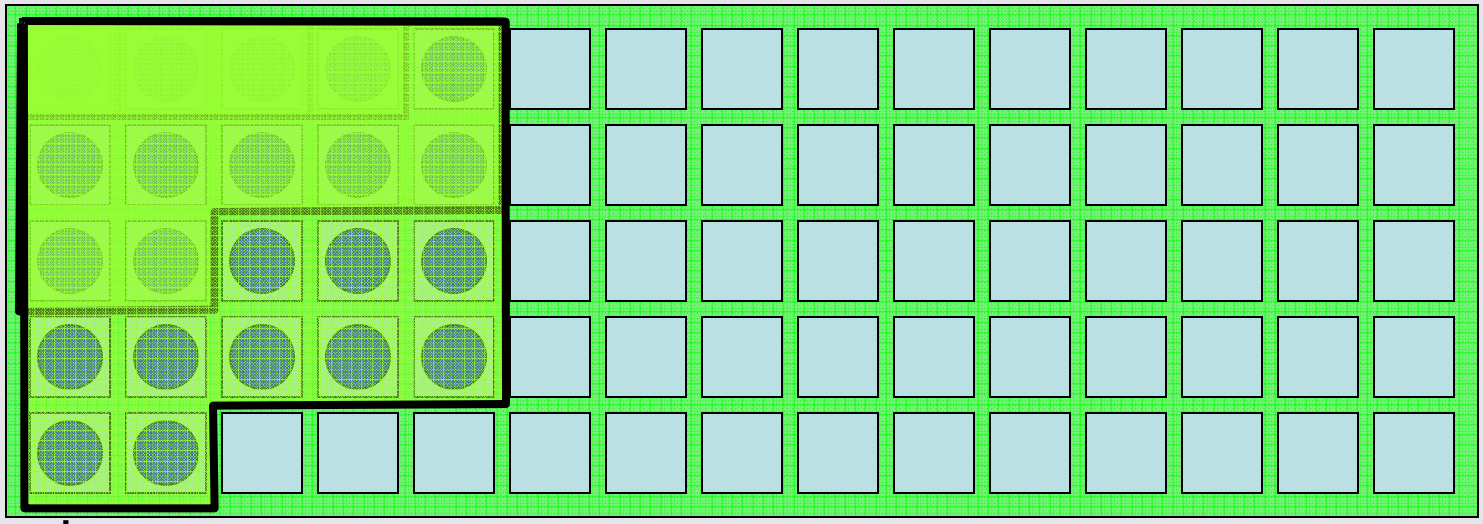


Building an Intranet

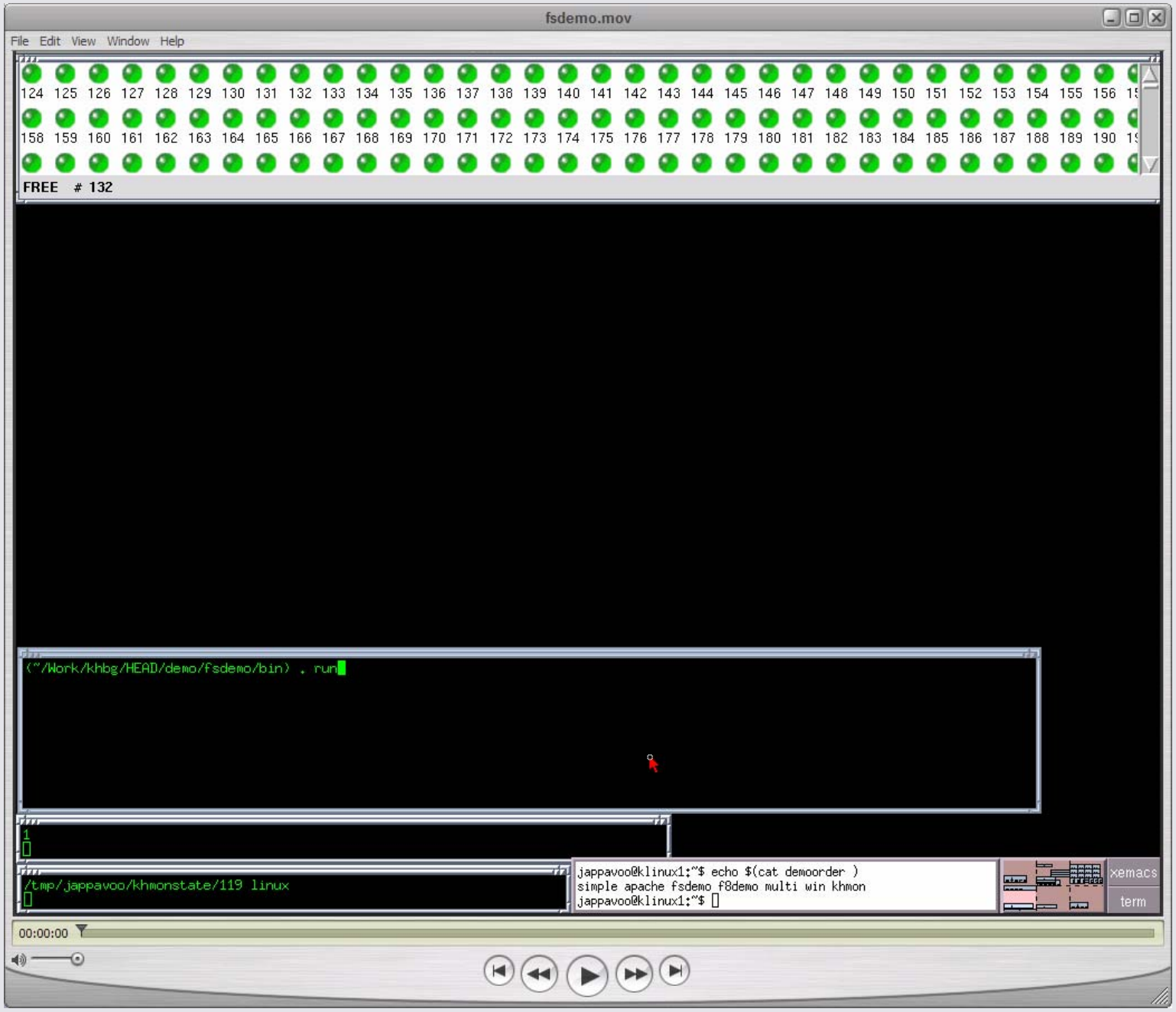
http://www.research.ibm.com/kittyhawk/movies/fsdemo.mov

A Simple Example on our Prototype

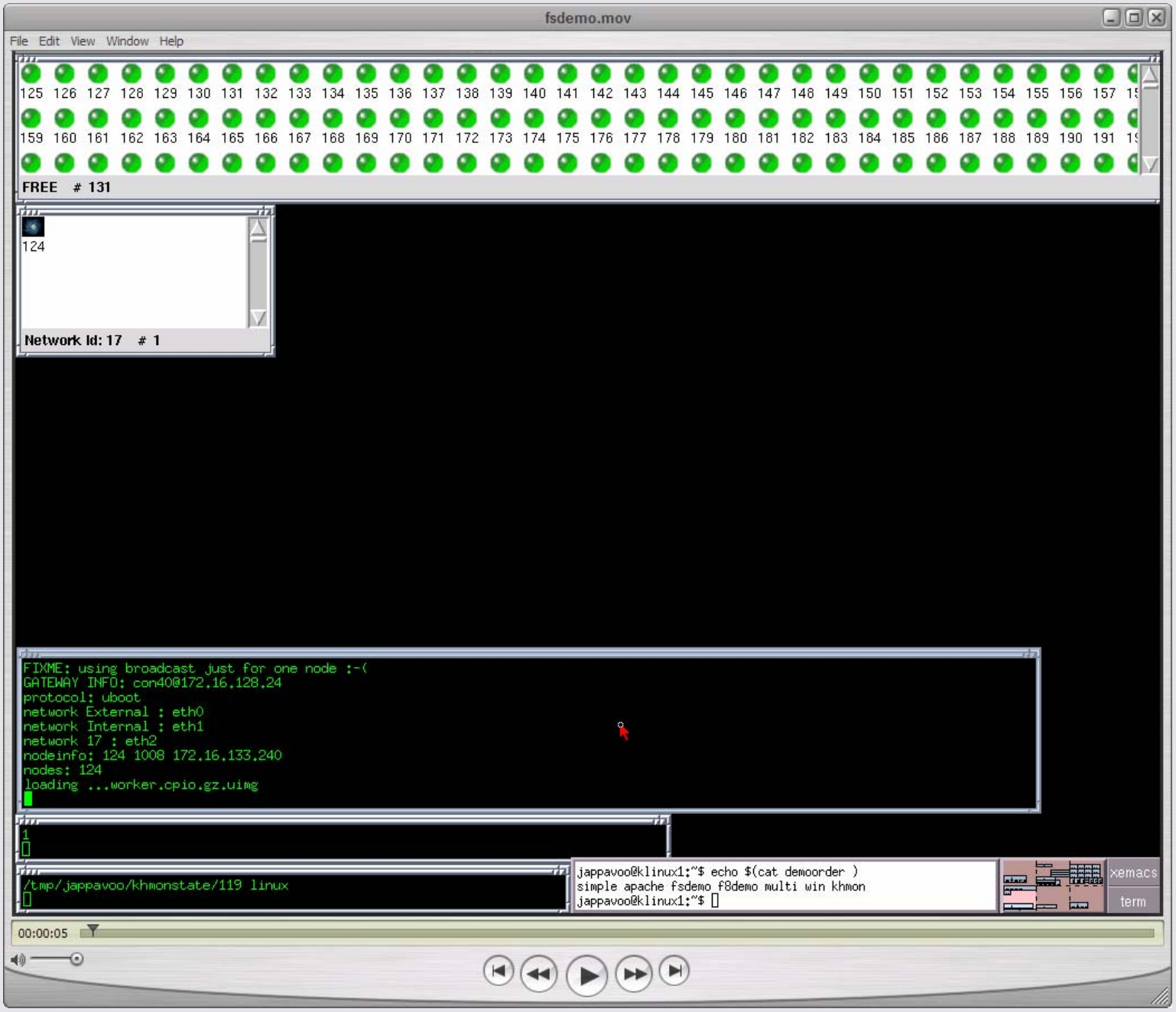




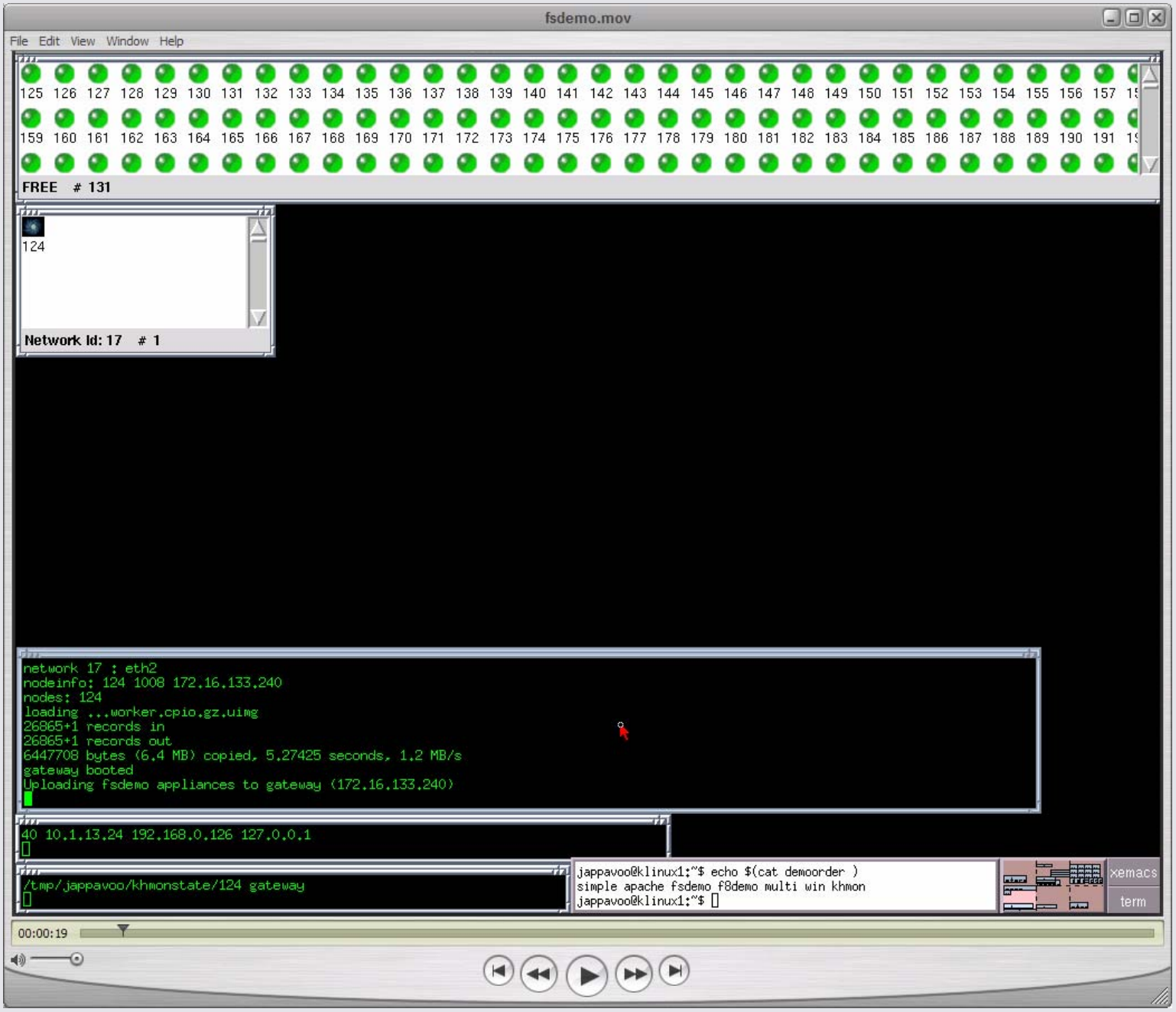
00:00



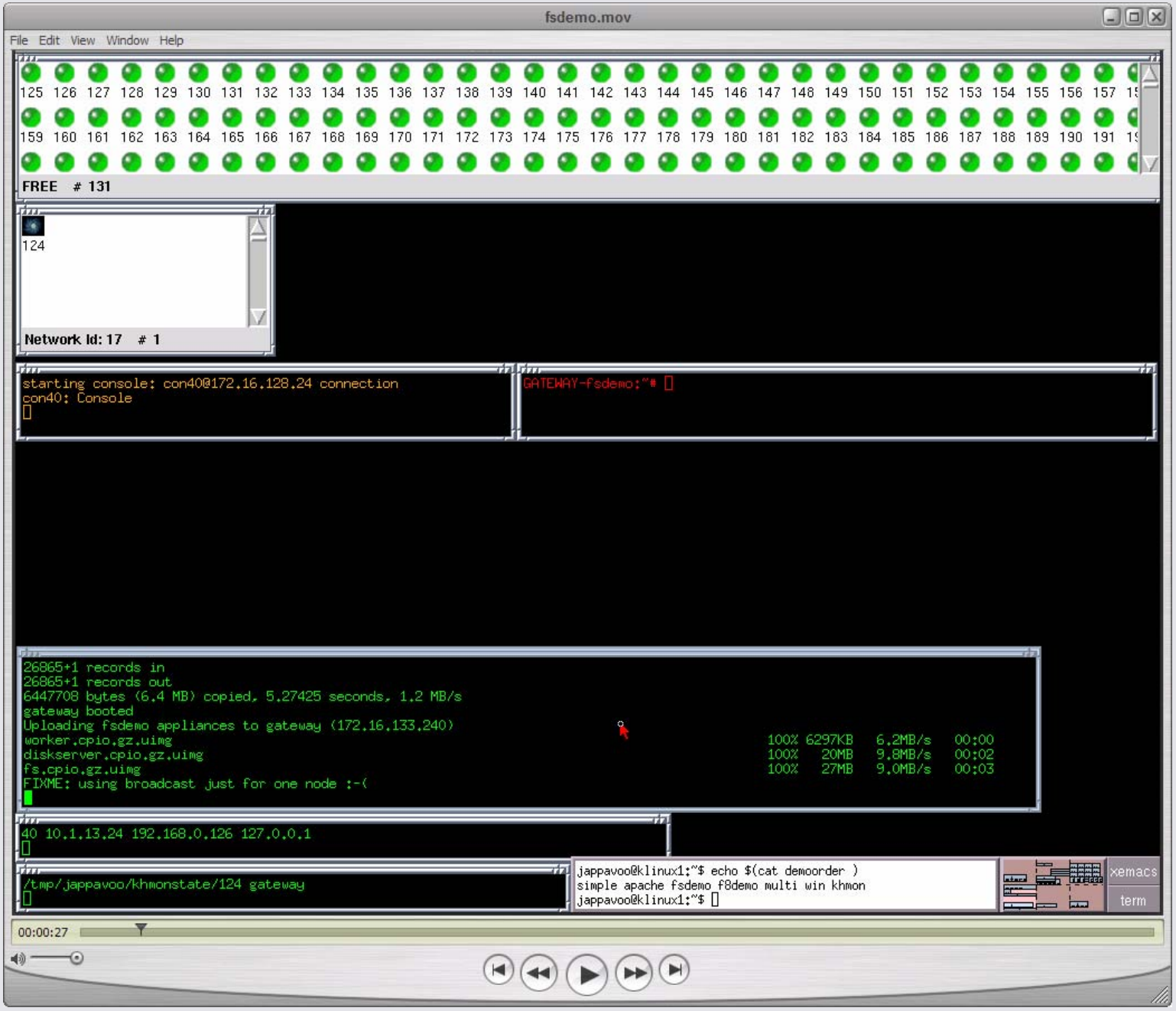
00:05



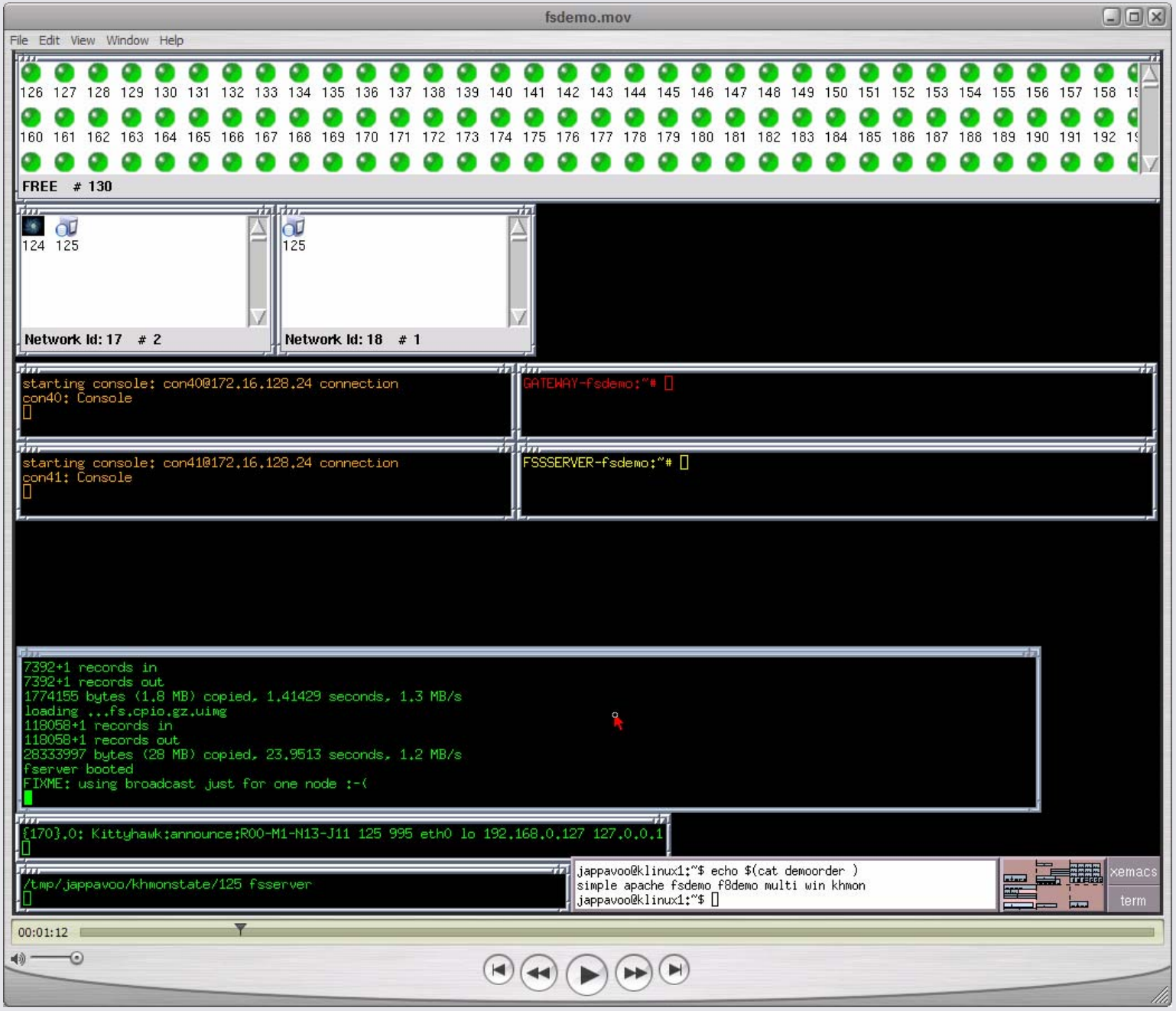
00:19



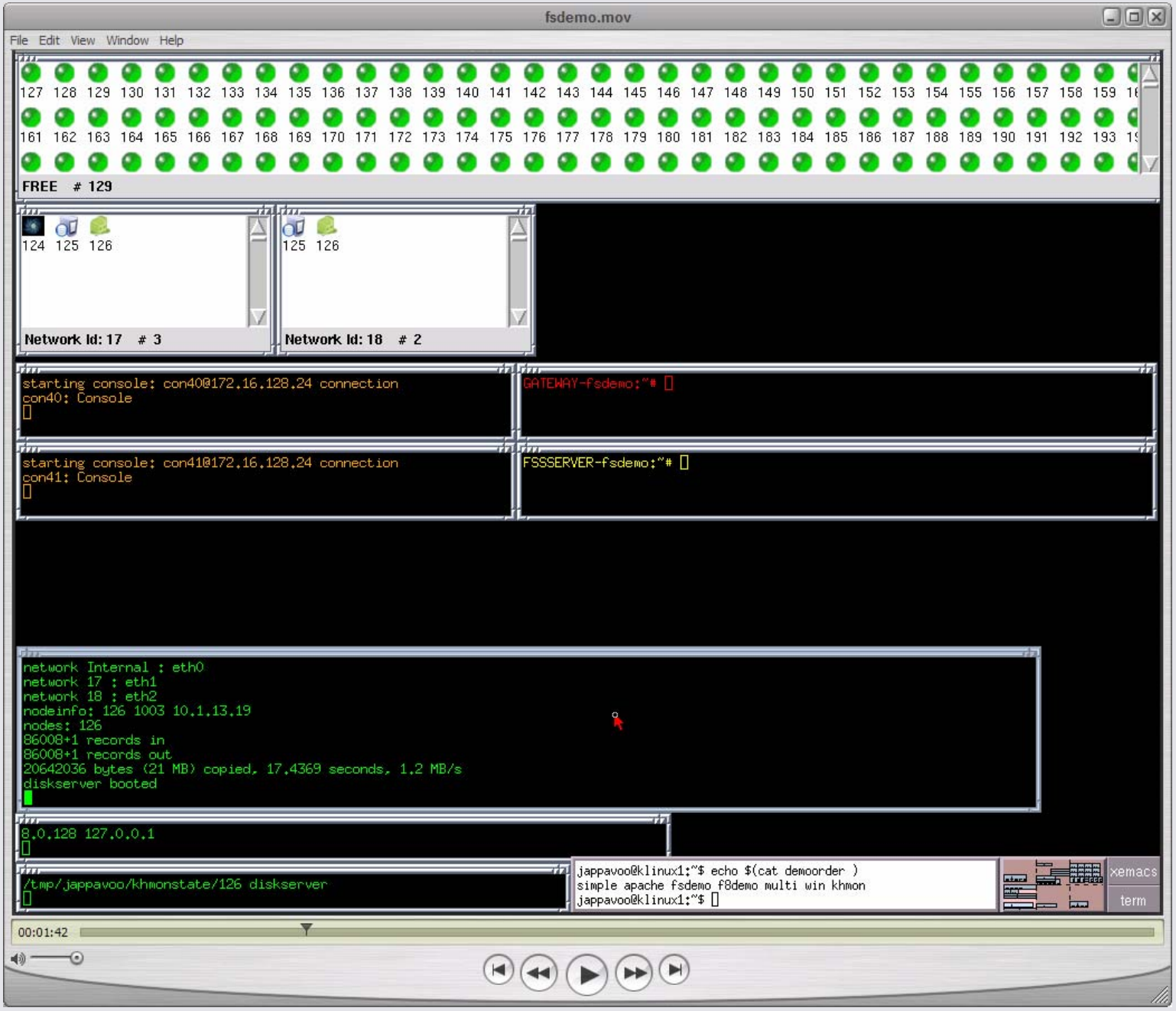
00:27



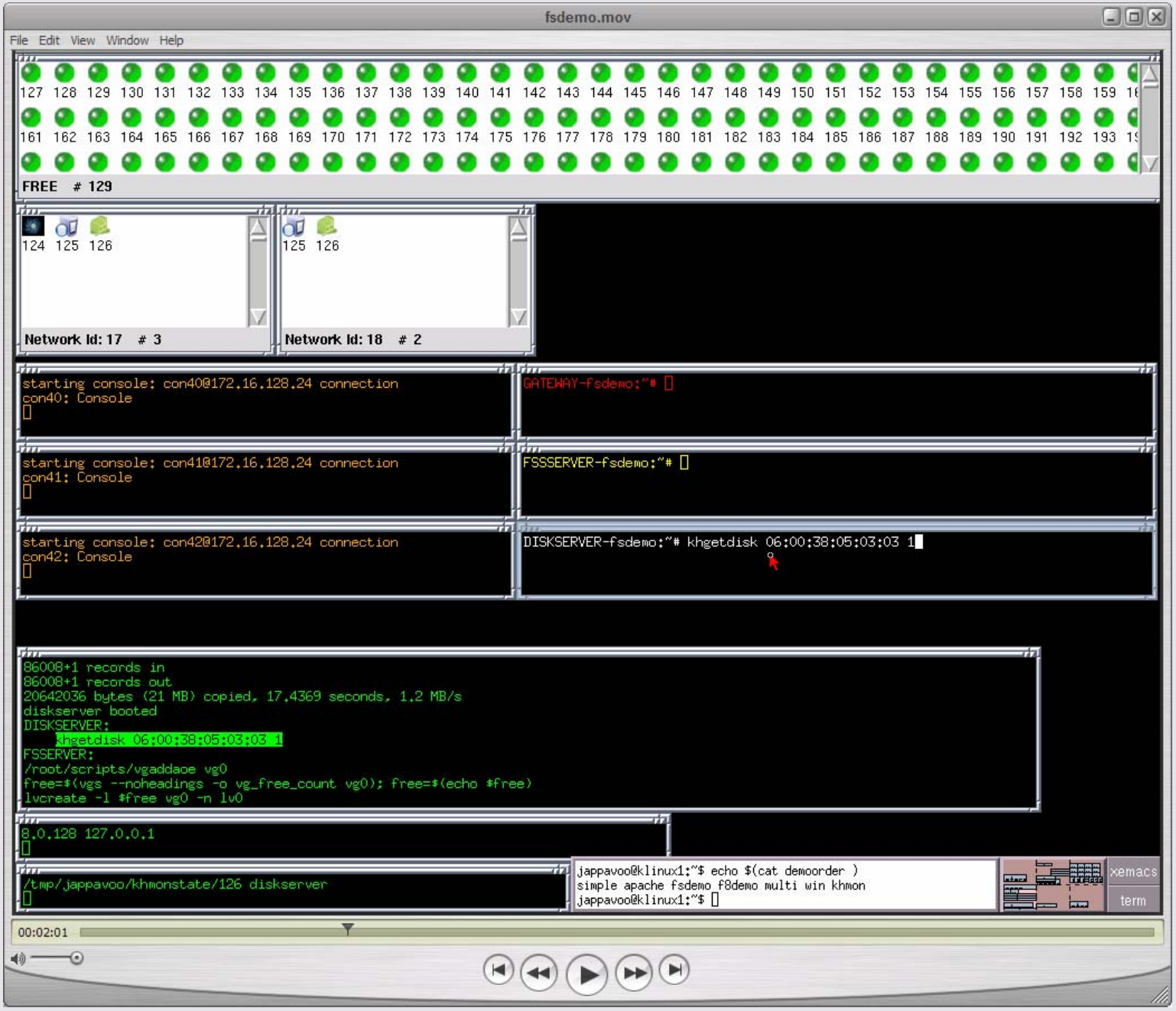
01:12



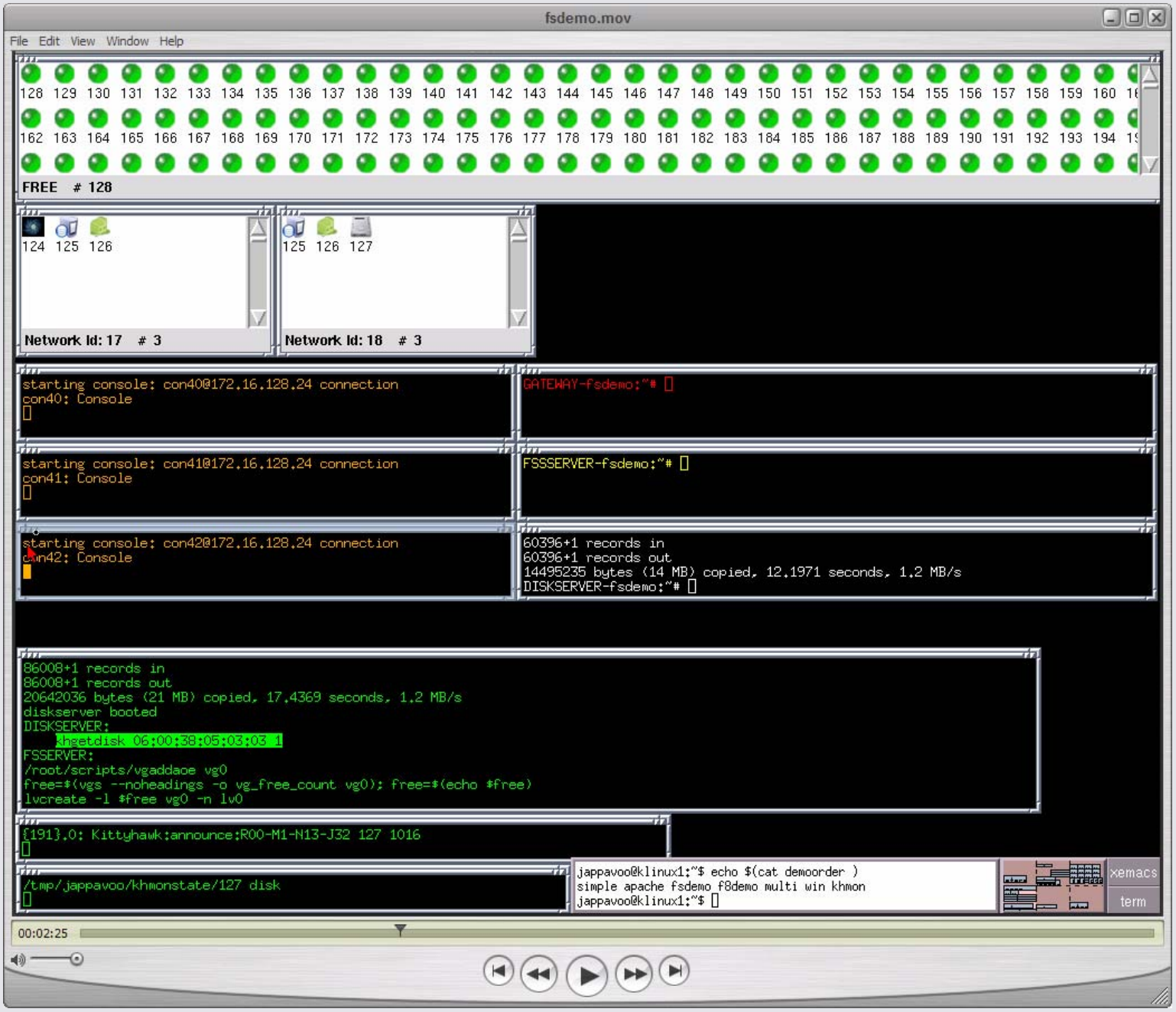
01:42



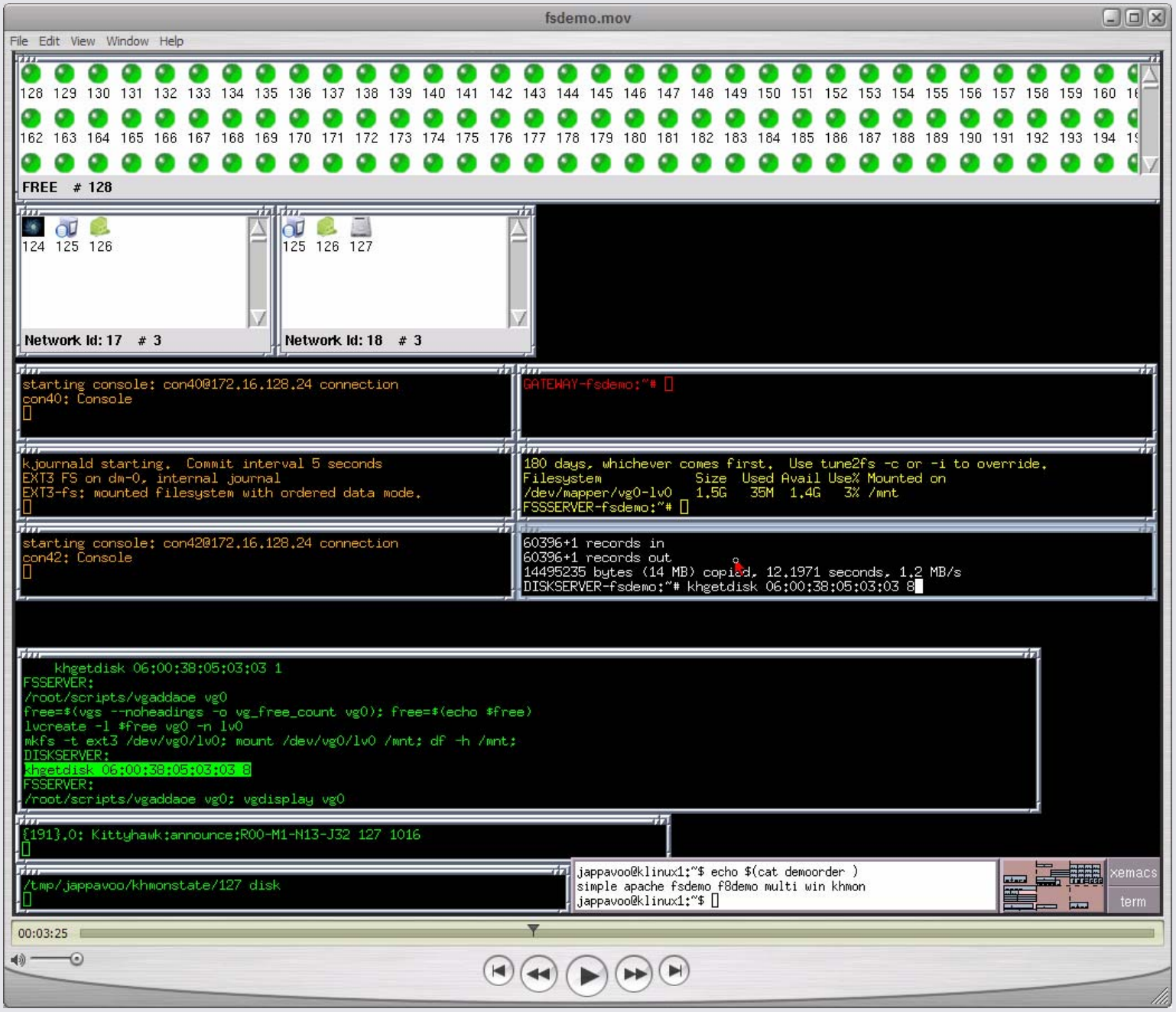
02:01



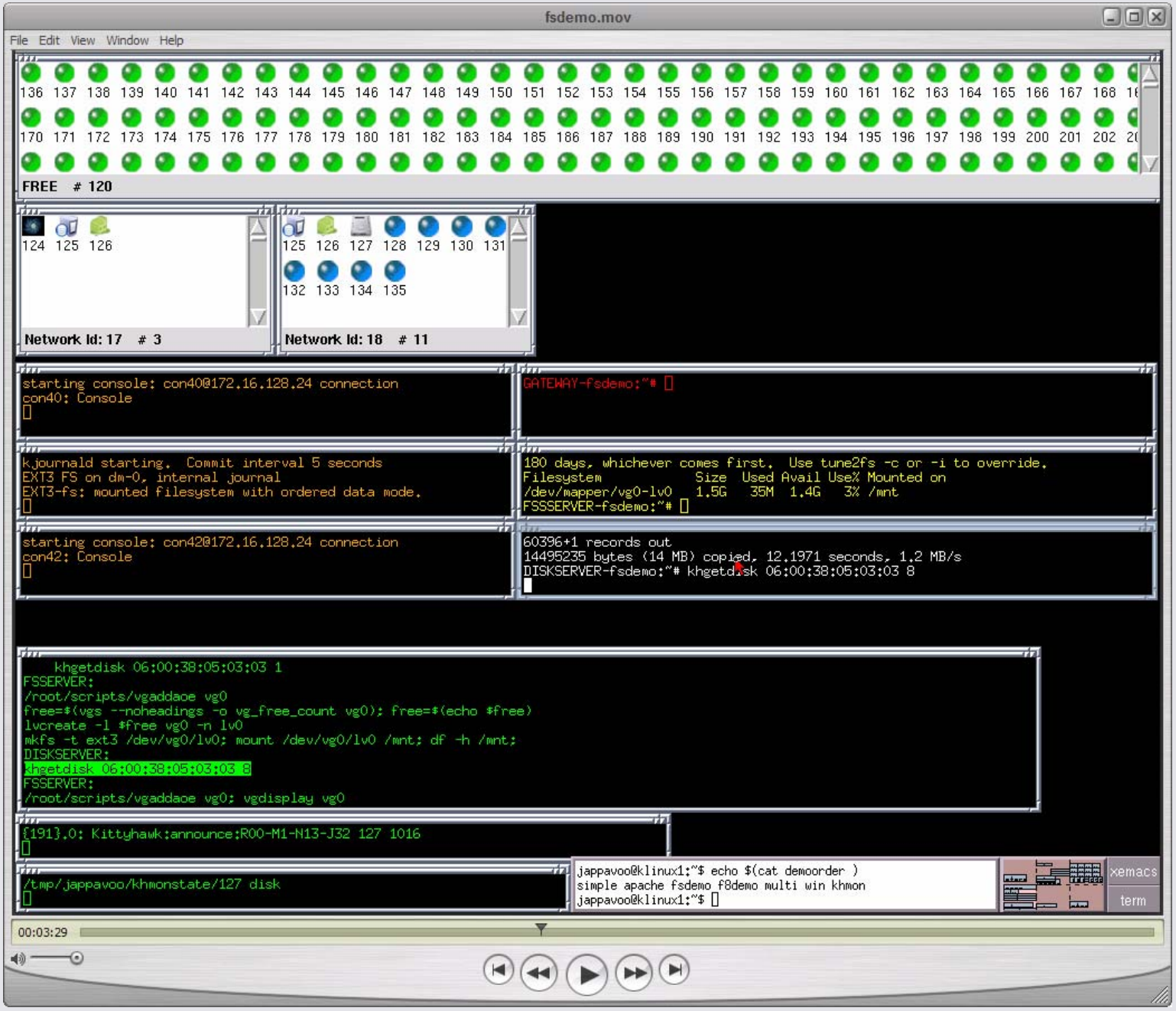
02:25



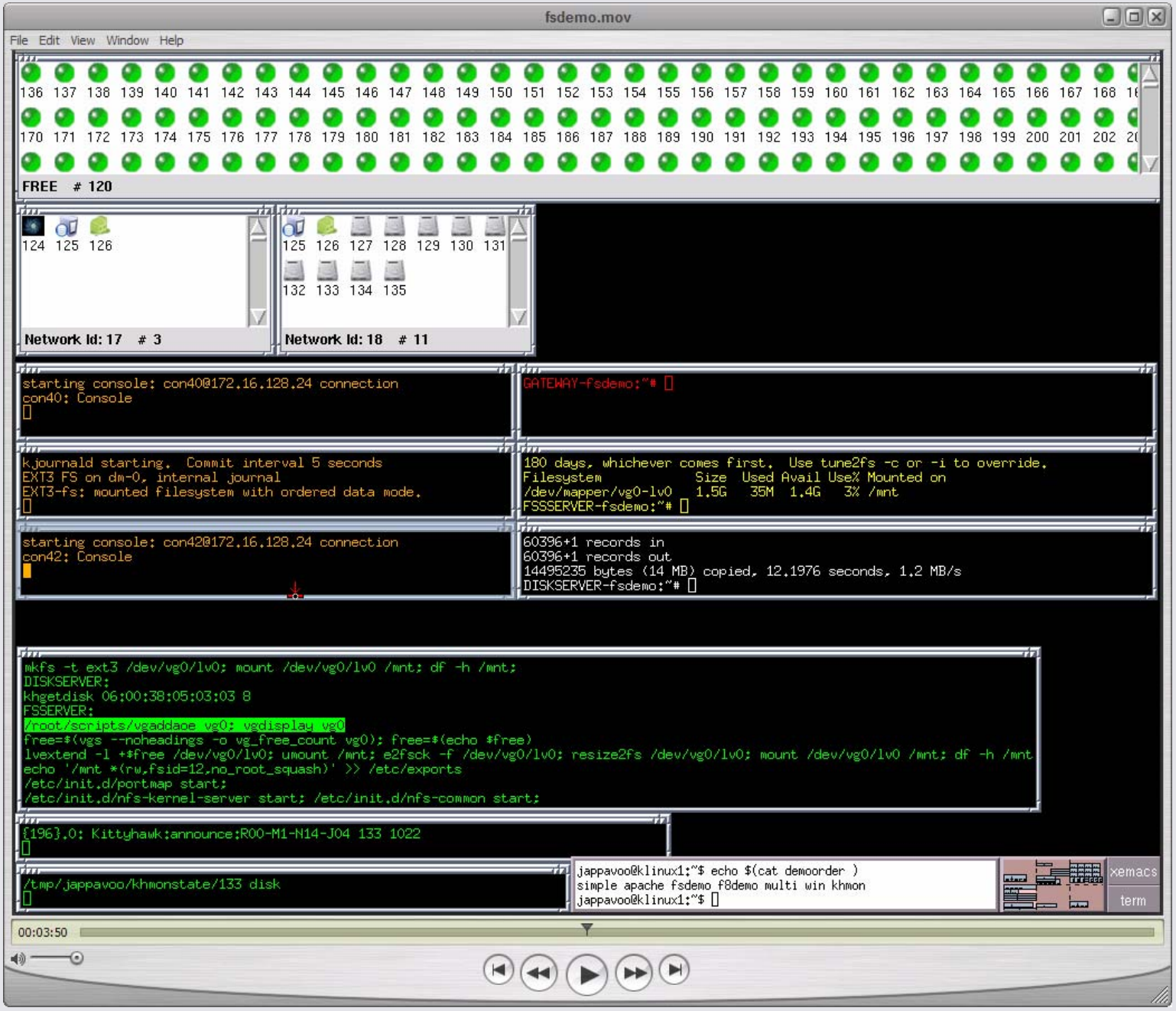
03:25



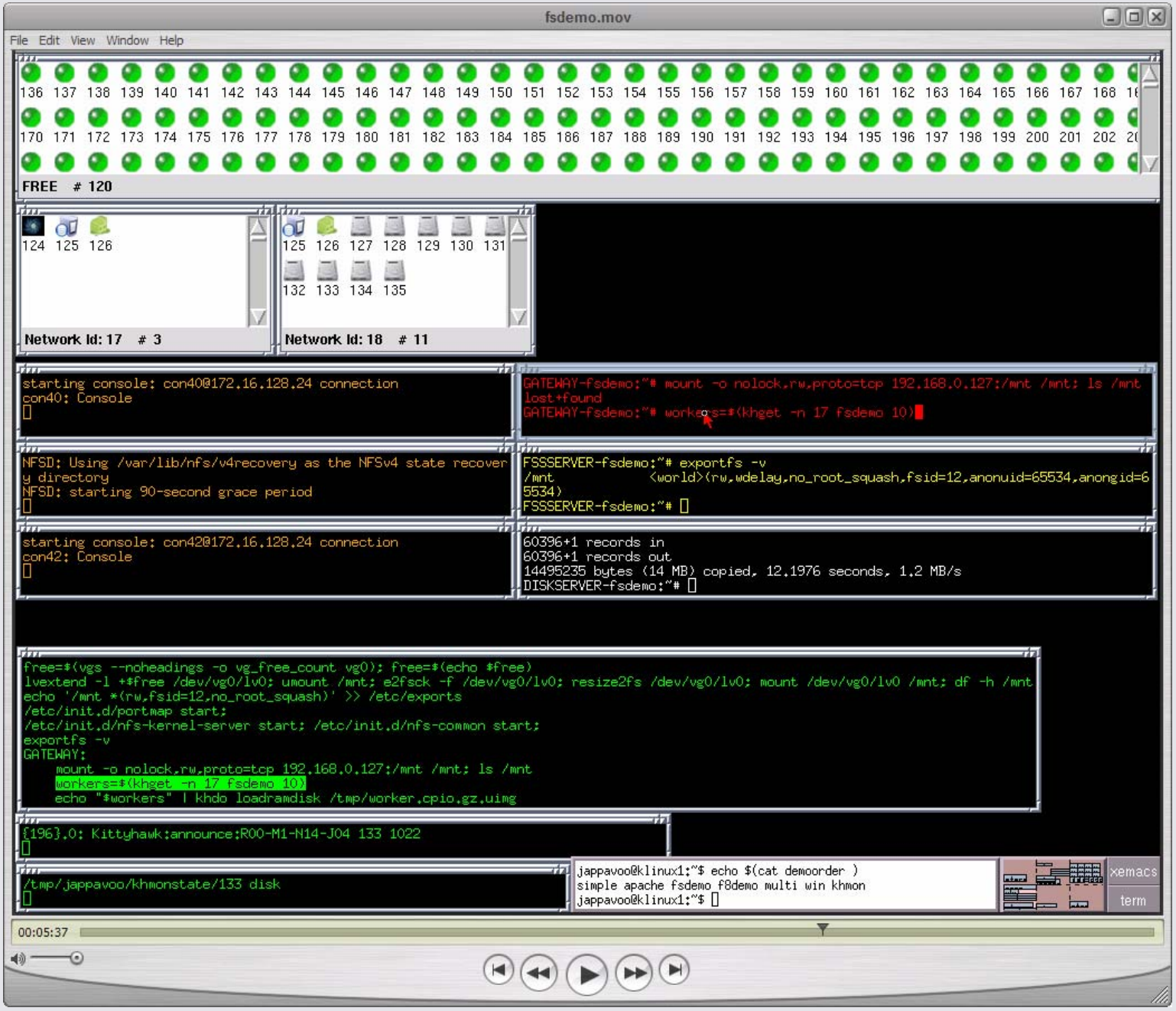
03:29



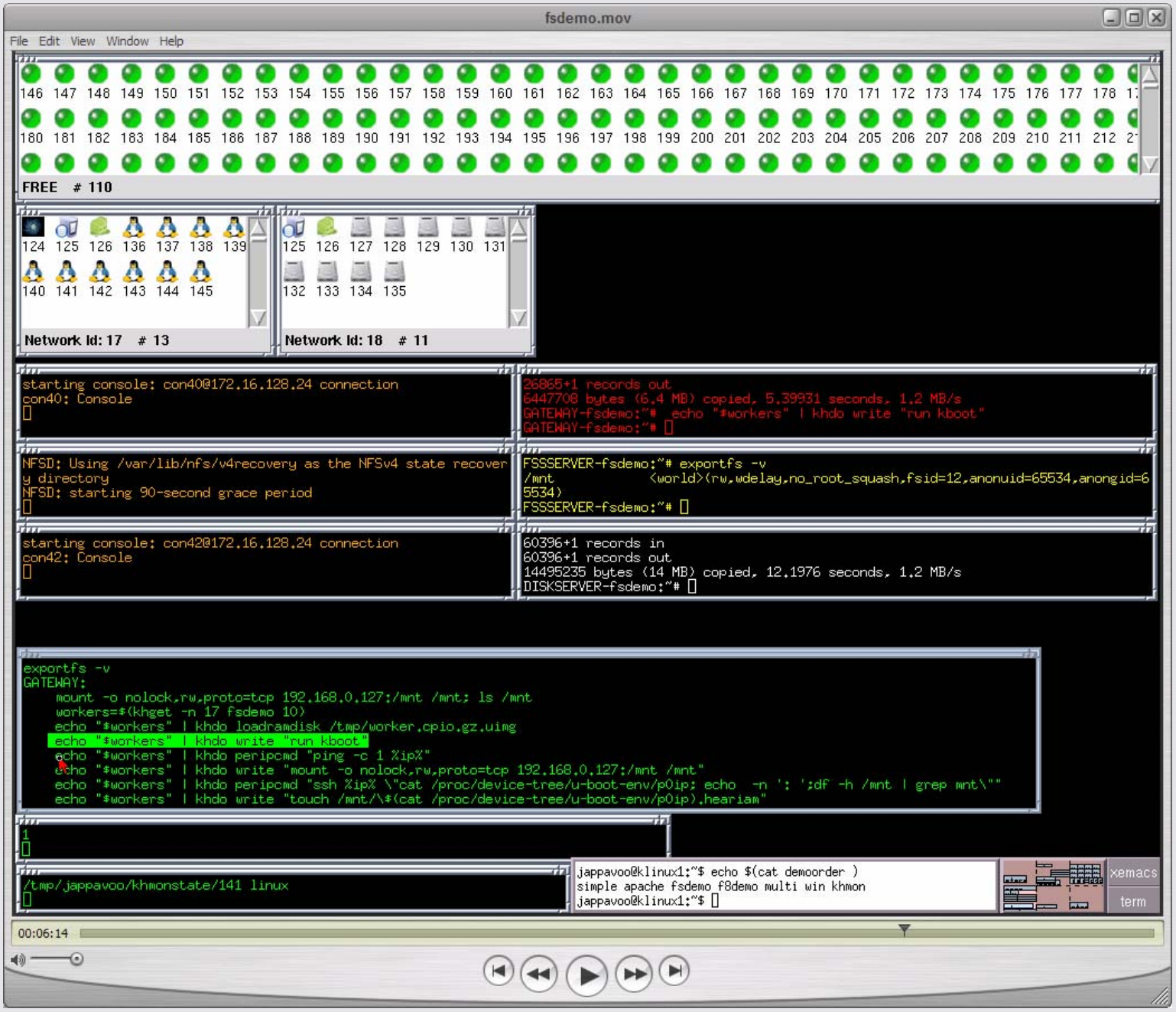
03:50



05:37



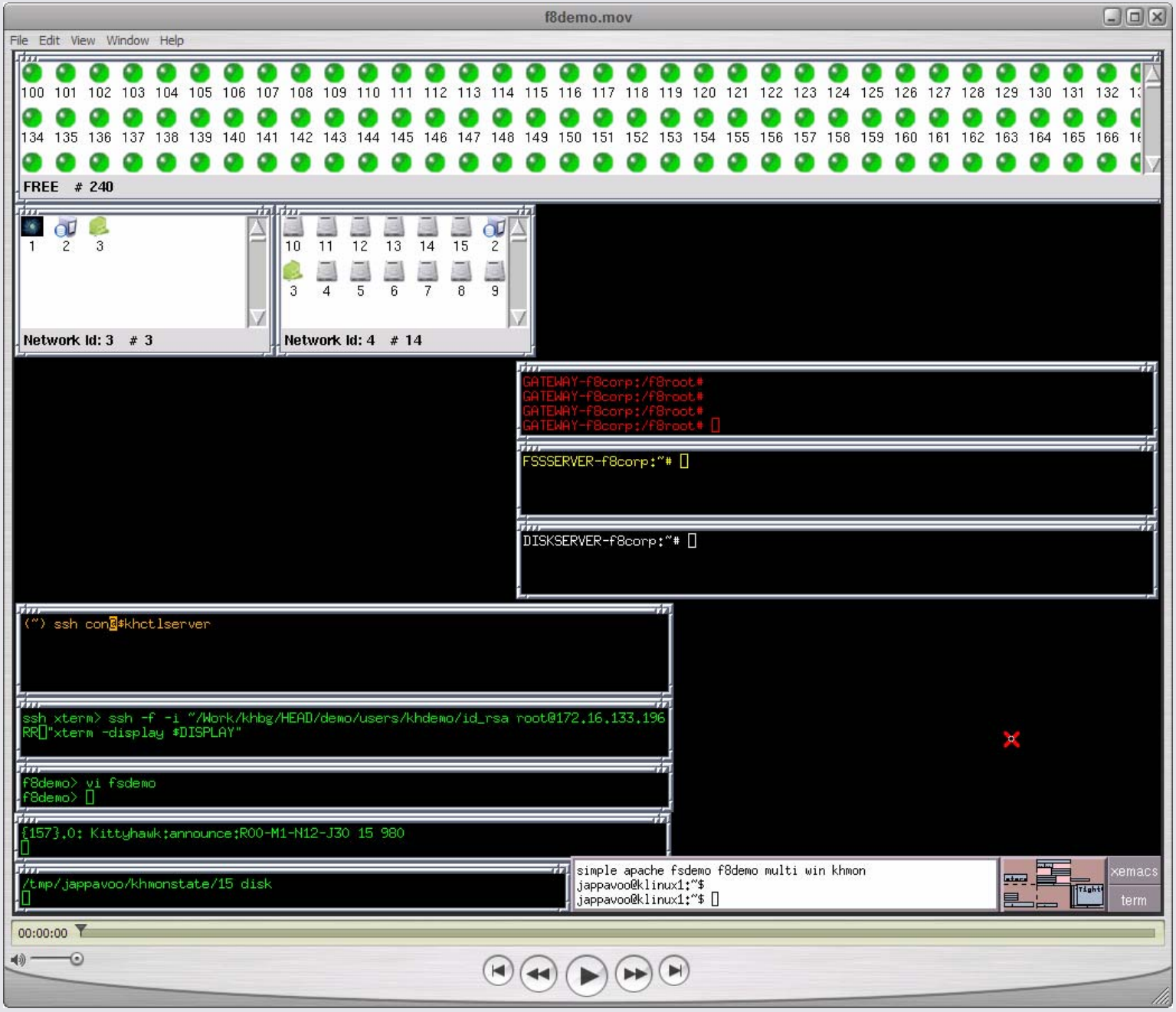
06:14



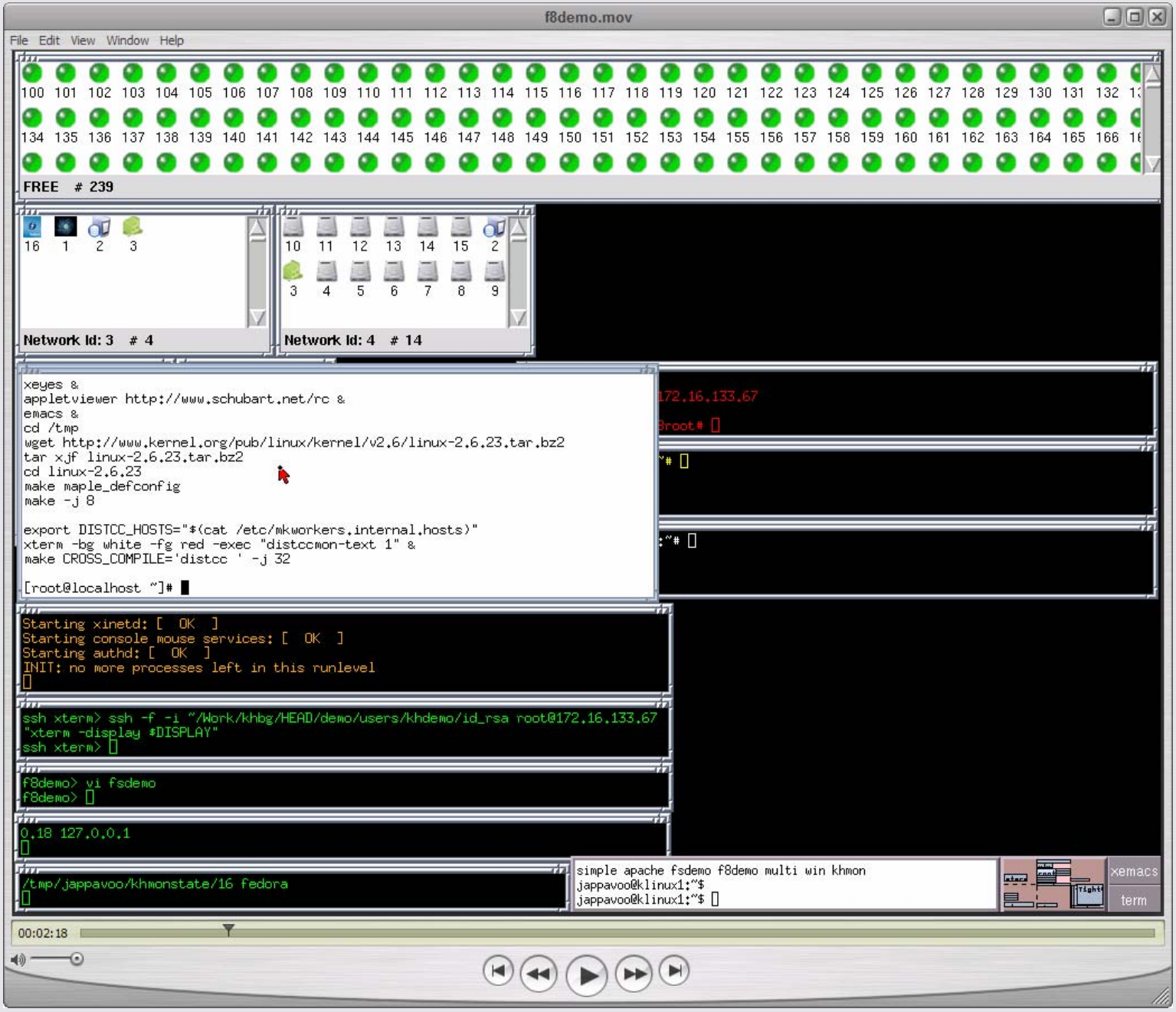
More than Appliances

http://www.research.ibm.com/kittyhawk/movies/f8demo.mov

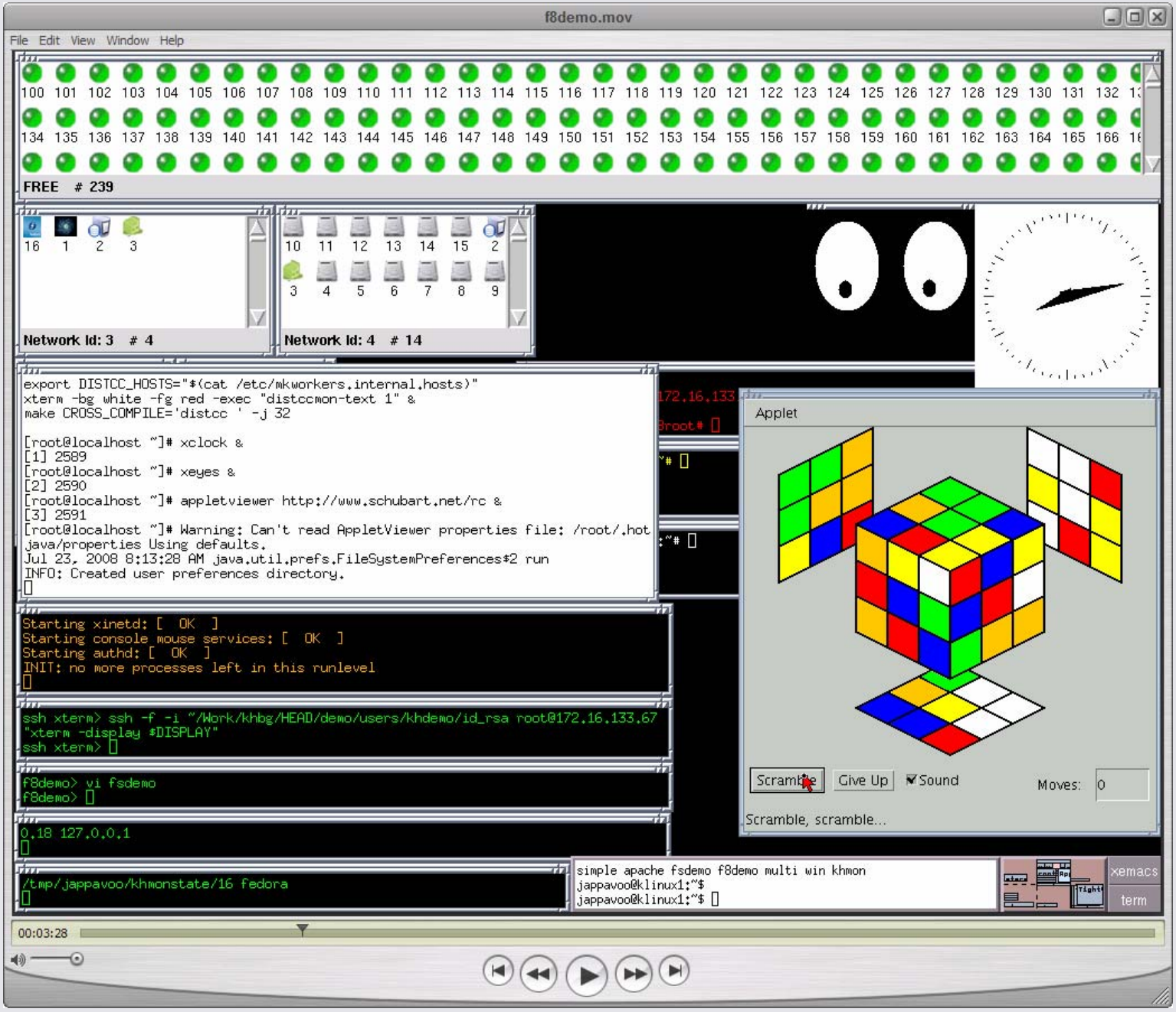
00:00



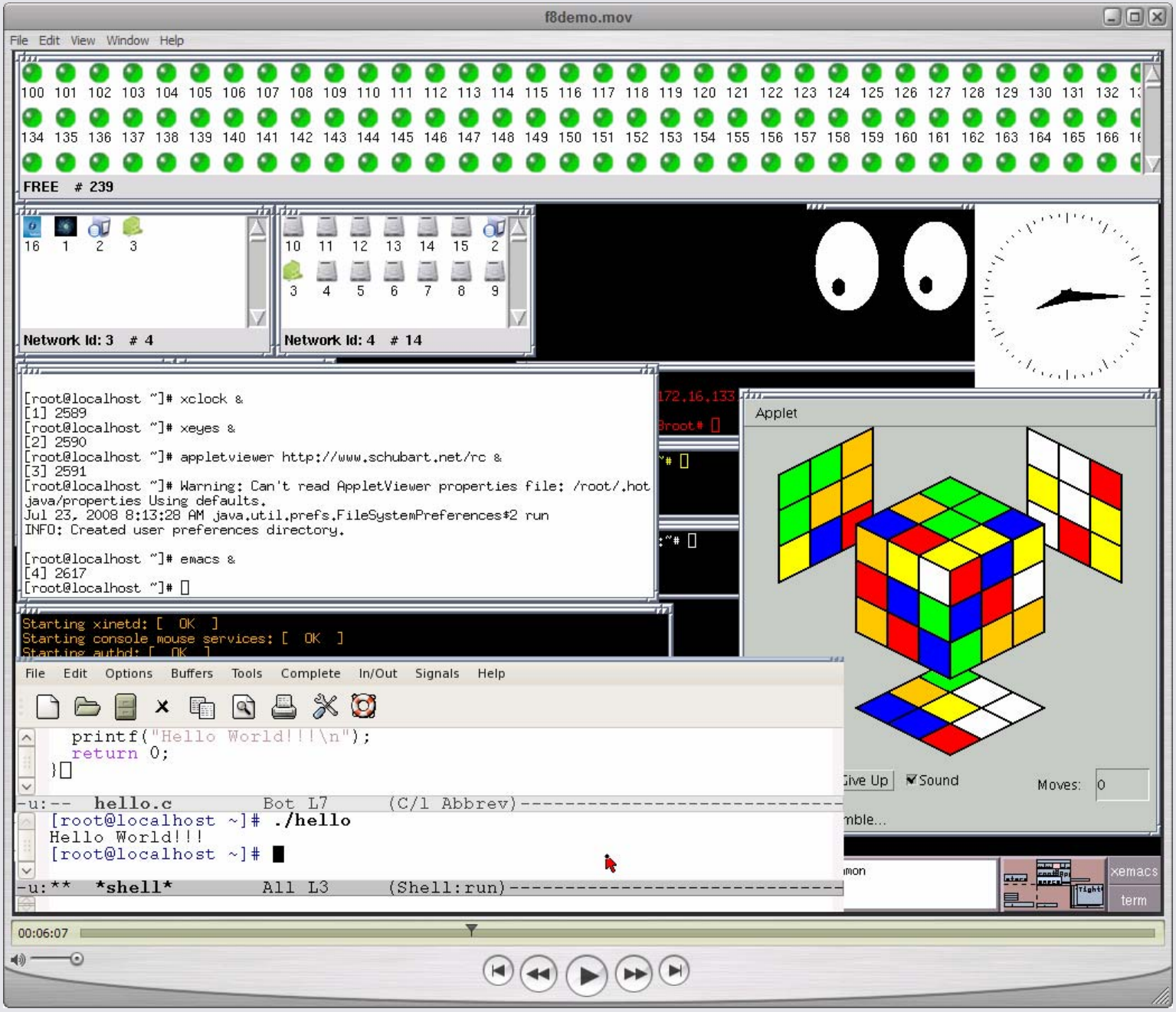
02:18



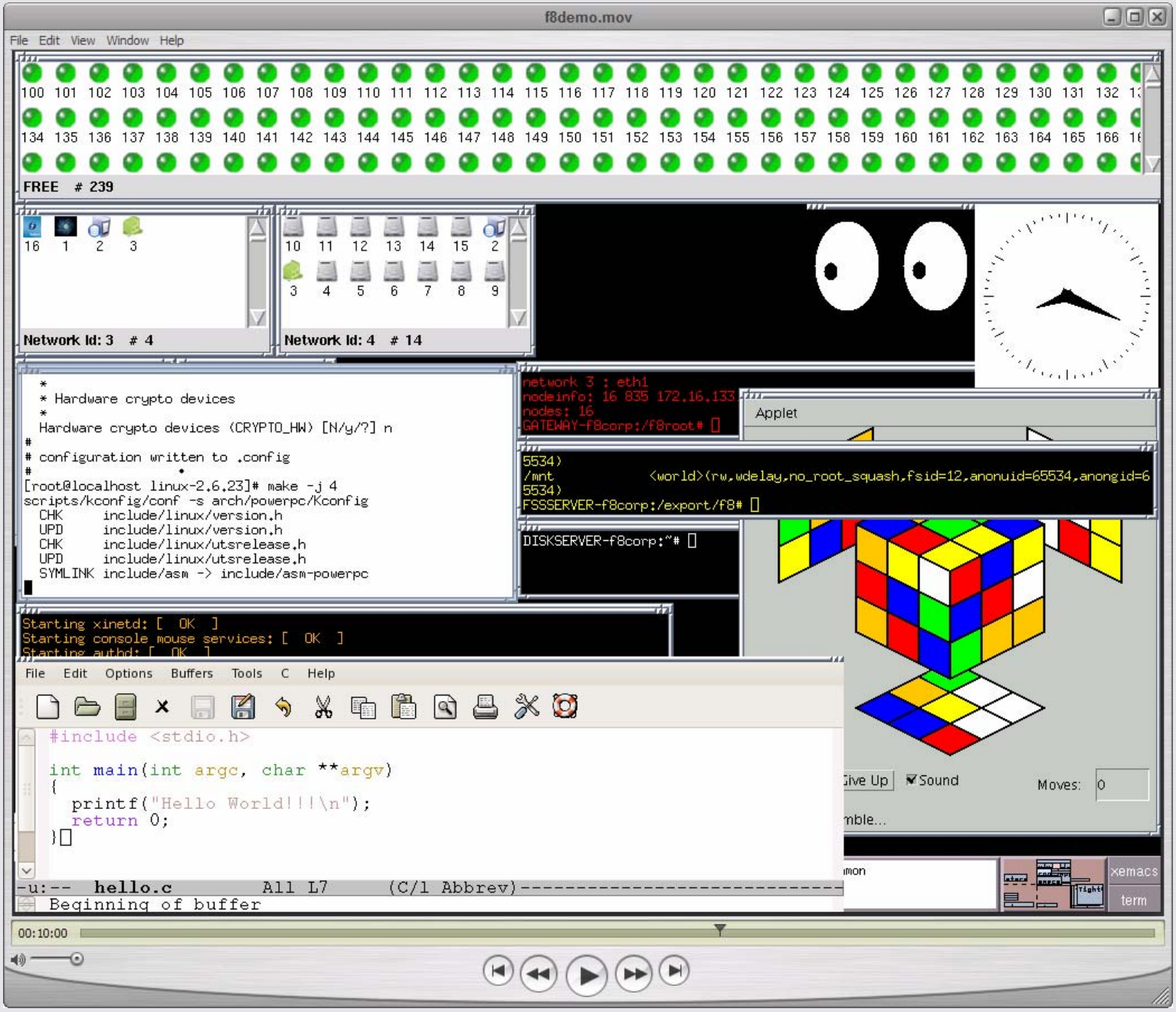
03:28



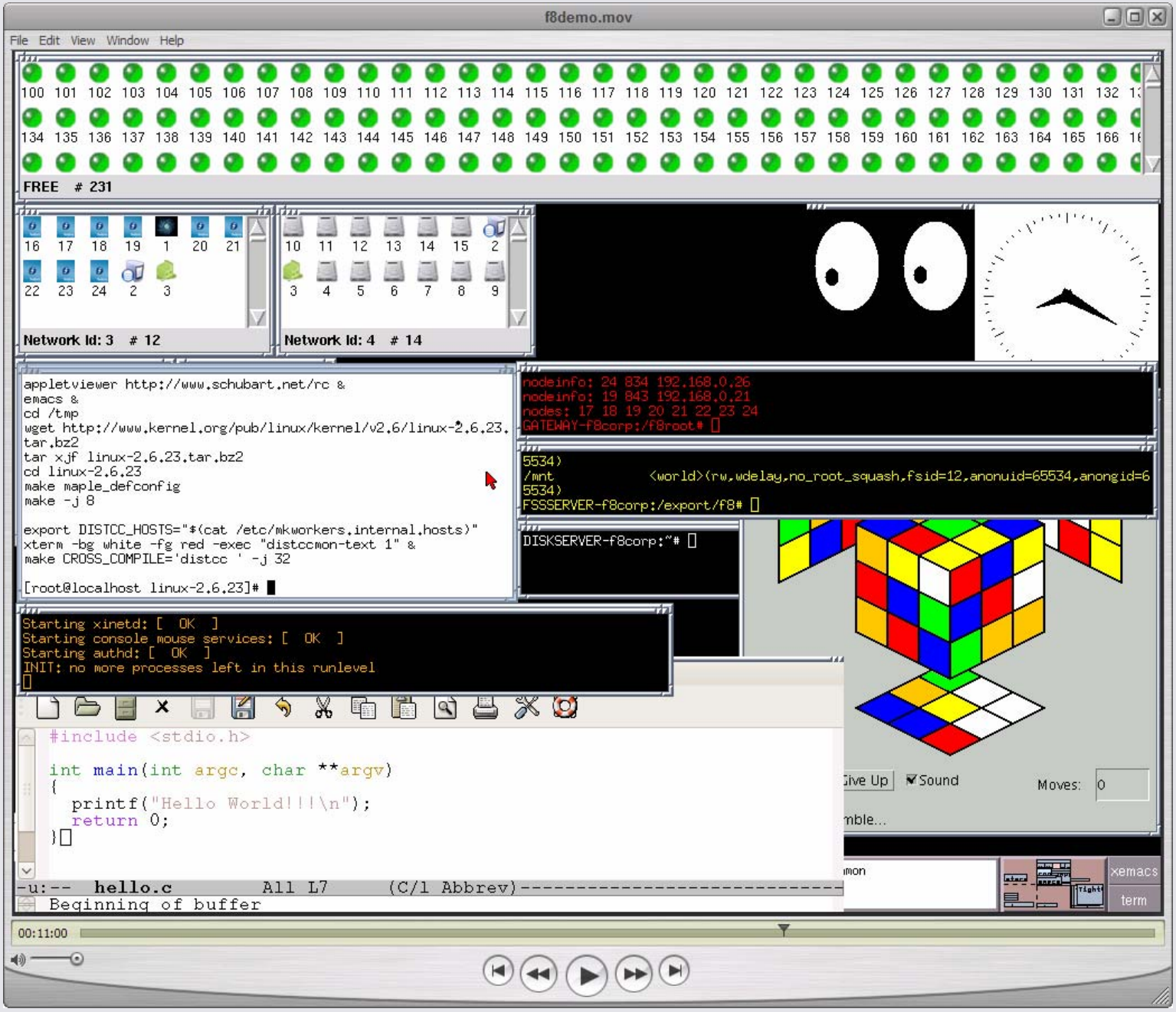
06:07



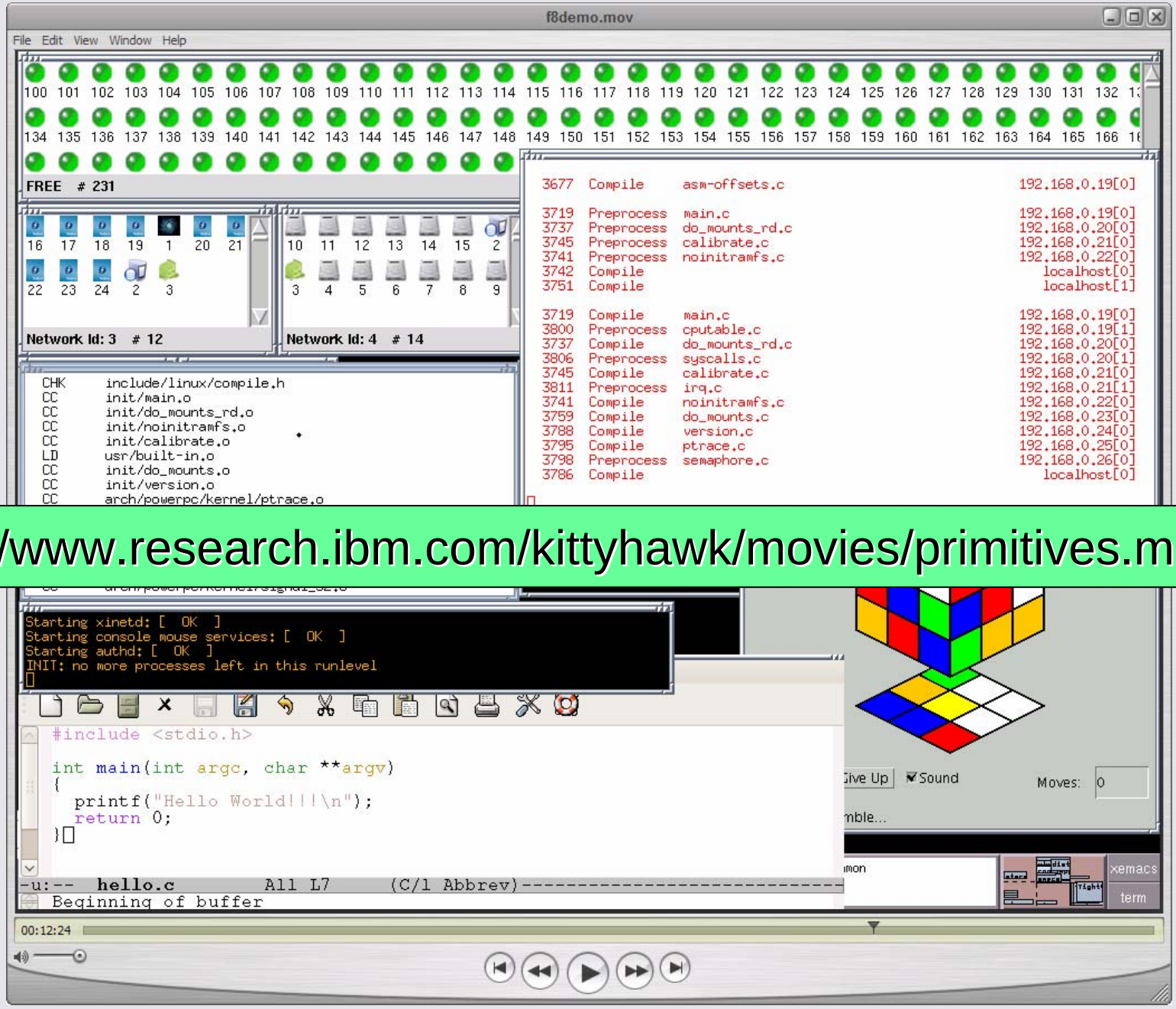
10:00



11:00



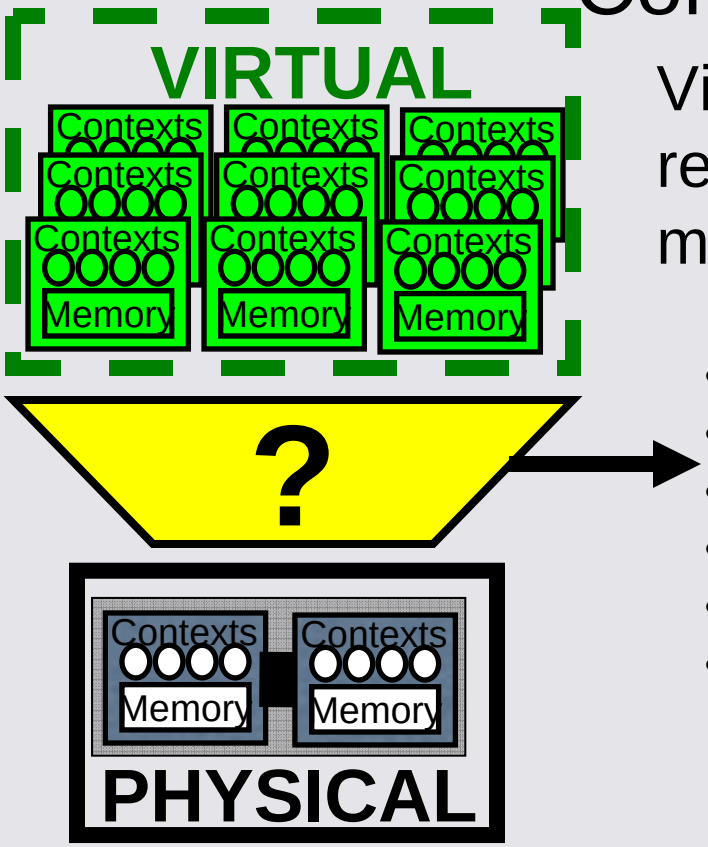
12:24



http://www.research.ibm.com/kittyhawk/movies/primitives.mov

Why Hardware-centric? When Virtual is Bad!

Heresy: Inherent virtualization runs counter to a fair system – A Virtual Computer is not a Computer!



Virtualization / time-sharing is real technology with complex mechanisms and policies.

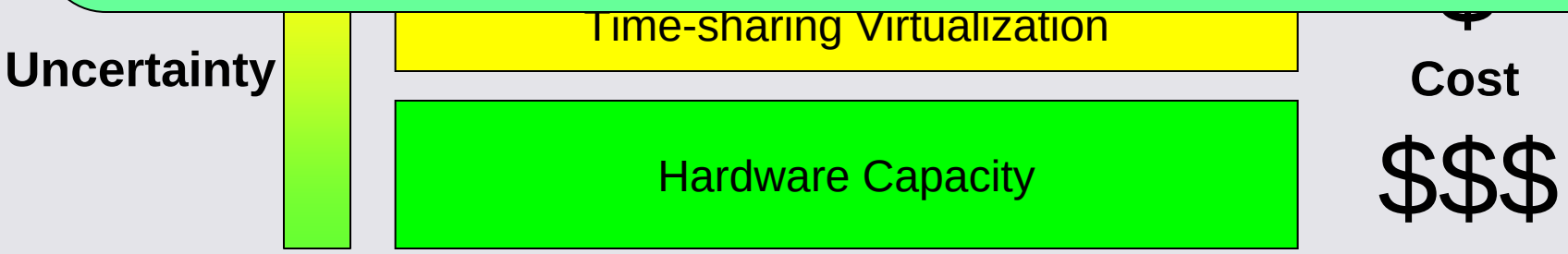
- Virtual Context → Physical Context?
- Virtual Memory → Physical Memory?
- Virtual I/O ←→ Real I/O?
- Security?
- Utilization?
- etc.

Virtualization is a service/derivative/future of the underlying resource. Its value should be isolated.

Uncertainty in Resources

"A user can either pick a level of service and pay whatever it takes to obtain it (giving him certainty with respect to service, uncertainty with respect to cost), or he can pick a price level and accept whatever level of service he can obtain

Fairness
through
Raw Hardware Access



Our Digital Future -- by Accident or Design?

"The past is but the beginning of a beginning, and all that is or has been is but the twilight of the dawn"

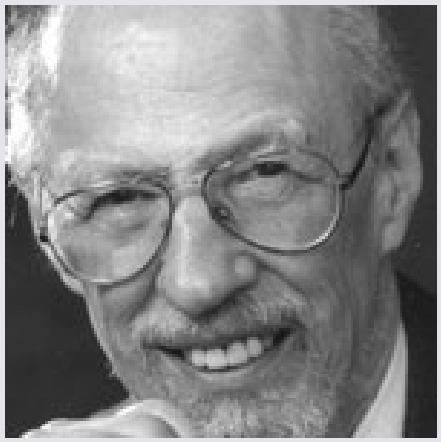
H.G. Wells, *"The Discovery of the Future"*,1901

Advice from the Pioneers

John McCarthy

Joseph Carl Robnett Licklider

Robert Fano



Fernando J. Corbató

"The very power of advanced computer systems makes them a serious threat to

"…multiple access systems operated as a utility will substantially extend the exploitation of information processing systems to the point of having significant social consequences."

F.J. Corbató and V.A. Vyssotsky, "Introduction and Overview of the Multics System", Proceedings of the 1965 Fall Joint Computer Conference.

Accessible Computing", 1965 Fall Joint Computer Conference.

Not If, but How?

- "The development of such a system is probably inevitable (unless it is forbidden by law)…"
- "The main danger to be avoided is the creation of services of limited scope…"
- "Another problem is to avoid monopolies; the intrinsic nature of the system permits any person who can write computer programs to compete with large organizations in inventing and offering imaginative services, but one can worry that the system might develop commercially in some way that would prevent that. In general we should try to develop information services in such a way as will enhance the individuality of its users."
- "… The major force that might tend to reduce competition is the exclusive possession of proprietary programs or files. Therefore it is desirable to separate the ownership of programs performing services from the ownership of the service bureaus themselves …"

John McCarthy, "The Home Information Terminal", Man and Computer. Proc. int. Conf., Bordeaux 1970, pp 48-57.

Summary

1. Large-scale communication-centric system through **Aggressive Integration.**
2. Fairness through **Raw Hardware**

Our Digital Future is at hand -- let's have some FUN!

Open questions abound; we don't have the answers; let's chat!

end, exploitation of computers for the benefit of society hinges upon two pivots: education, and **responsible considered action by those of the technical community able to exert some influence**."

E.E. David Jr and R.M. Fano, "Some Thoughts About the Social Implications of Accessible Computing", 1965 Fall Joint Computer Conference.