CS162
Operating Systems and Systems Programming
Lecture 26

Protection and Security II, ManyCore Operating Systems

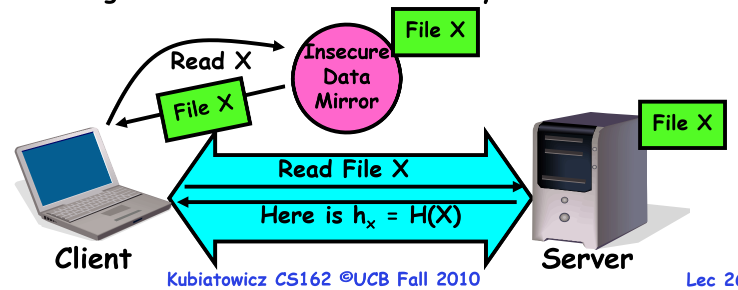
December 1st, 2010

Prof. John Kubiatowicz

http://inst.eecs.berkeley.edu/~cs162

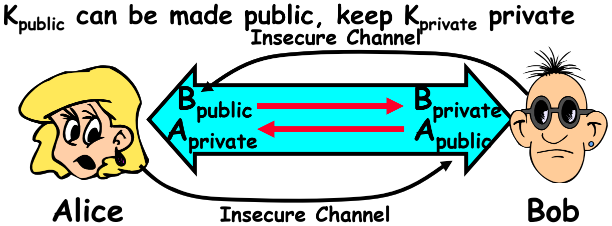
Review: Use of Hash Functions

- Several Standard Hash Functions:
  - MD5: 128-bit output
  - SHA-1: 160-bit output, SHA-256: 256-bit output
- Can we use hashing to securely reduce load on server?
  - Yes. Use a series of insecure mirror servers (caches)
  - First, ask server for digest of desired file
    - » Use secure channel with server
  - Then ask mirror server for file
    - » Can be insecure channel
    - » Check digest of result and catch faulty or malicious mirrors



Review: Public Key Encryption Details

- Idea: K_{public} can be made public, keep $K_{private}$ private



- Gives message privacy (restricted receiver):
  - Public keys can be acquired by anyone/used by anyone
  - Only person with private key can decrypt message
- What about authentication?
  - Alice→Bob: [(I'm Alice)^{Aprivate} Rest of message]^{Bpublic}
  - Provides restricted sender and receiver
- Suppose we want X to sign message M?
  - Use private key to encrypt the digest, i.e. $H(M)^{Xprivate}$
  - Send both M and its signature: [M,$H(M)^{Xprivate}$]
  - Now, anyone can verify that M was signed by X
    - » Simply decrypt the digest with X_{public}
    - » Verify that result matches H(M)

Goals for Today

- Use of Cryptographic Mechanisms
- Distributed Authorization/Remote Storage
- Worms and Viruses
- ManyCore operating systems

Note: Some slides and/or pictures in the following are adapted from slides ©2005 Silberschatz, Galvin, and Gagne. Also, slides on Taint Tracking adapted from Nickolai Zeldovich

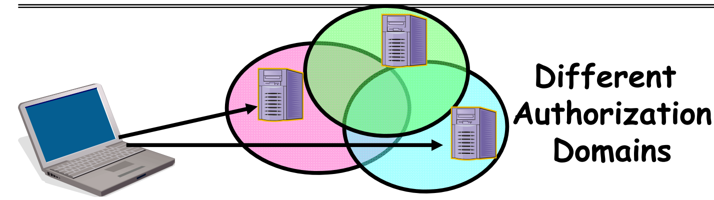
Recall: Authorization: Who Can Do What?

- How do we decide who is authorized to do actions in the system?
- Access Control Matrix: contains all permissions in the system
  - Resources across top
    - » Files, Devices, etc…
  - Domains in columns
    - » A domain might be a user or a group of permissions
    - » E.g. above: User D_3 can read F_2 or execute F_3
  - In practice, table would be huge and sparse!

domain \ object	F_1	F_2	F_3	printer
D_1	read		read	
D_2				print
D_3		read	execute	
D_4	read write		read write	

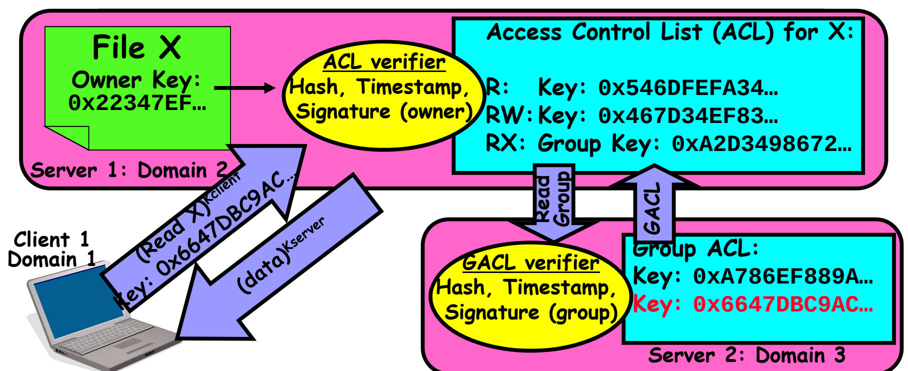
- Two approaches to implementation
  - Access Control Lists: store permissions with each object
    - » Still might be lots of users!
    - » UNIX limits each file to: r,w,x for owner, group, world
    - » More recent systems allow definition of groups of users and permissions for each group
  - Capability List: each process tracks objects has permission to touch
    - » Popular in the past, idea out of favor today
    - » Consider page table: Each process has list of pages it has access to, not each page has list of processes …

How to perform Authorization for Distributed Systems?



- Issues: Are all user names in world unique?
  - No! They only have small number of characters
    - » kubi@mit.edu → kubitron@lcs.mit.edu → kubitron@cs.berkeley.edu
    - » However, someone thought their friend was kubi@mit.edu and I got very private email intended for someone else…
  - Need something better, more unique to identify person
- Suppose want to connect with any server at any time?
  - Need an account on every machine! (possibly with different user name for each account)
  - OR: Need to use something more universal as identity
    - » Public Keys! (Called "Principles")
    - » People *are* their public keys

Distributed Access Control



- Distributed Access Control List (ACL)
  - Contains list of attributes (Read, Write, Execute, etc) with attached identities (Here, we show public keys)
    - » ACLs signed by owner of file, only changeable by owner
    - » Group lists signed by group key
  - ACLs can be on different servers than data
    - » Signatures allow us to validate them
    - » ACLs could even be stored separately from verifiers

Analysis of Previous Scheme

- Positive Points:
  - Identities checked via signatures and public keys
    - » Client can't generate request for data unless they have private key to go with their public identity
    - » Server won't use ACLs not properly signed by owner of file
  - No problems with multiple domains, since identities designed to be cross-domain (public keys domain neutral)
- Revocation:
  - What if someone steals your private key?
    - » Need to walk through all ACLs with your key and change…!
    - » This is very expensive
  - Better to have unique string identifying you that people place into ACLs
    - » Then, ask Certificate Authority to give you a certificate matching unique string to your current public key
    - » Client Request: (request + unique ID)$^{C_{private}}$; give server certificate if they ask for it.
    - » Key compromise⇒must distribute "certificate revocation", since can't wait for previous certificate to expire.
  - What if you remove someone from ACL of a given file?
    - » If server caches old ACL, then person retains access!
    - » Here, cache inconsistency leads to security violations!

Analysis Continued

- Who signs the data?
  - Or: How does client know they are getting valid data?
  - Signed by server?
    - » What if server compromised? Should client trust server?
  - Signed by owner of file?
    - » Better, but now only owner can update file!
    - » Pretty inconvenient!
  - Signed by group of servers that accepted latest update?
    - » If must have signatures from all servers ⇒ Safe, but one bad server can prevent update from happening
    - » Instead: ask for a threshold number of signatures
    - » Byzantine agreement can help here
- How do you know that data is up-to-date?
  - Valid signature only means data is valid older version
  - Freshness attack:
    - » Malicious server returns old data instead of recent data
    - » Problem with both ACLs and data
    - » E.g.: you just got a raise, but enemy breaks into a server and prevents payroll from seeing latest version of update
  - Hard problem
    - » Needs to be fixed by invalidating old copies or having a trusted group of servers (Byzantine Agrement?)

Administrivia

- Optional Lecture on Monday at normal time and place
  - Topics still TBA, but it will be good! ☺
- Final Exam
  - Thursday 12/16, 8:00AM-11:00AM, 10 Evans Hall
  - All material from the course
    - » With slightly more focus on second half
  - Two sheets of notes, both sides
  - Will need dumb calculator
- Should be working on Project 4
  - Final Project due on Tuesday 12/7
- I will have office hours next week at normal time
  - M/W 2:30-3:30
  - Feel free to come by to talk about whatever
- Need to get any regrade requests in by next Friday
  - i.e. Projects 1-3
  - Will consider Project 4 issues up until final

Involuntary Installation

- What about software loaded without your consent?
  - Macros attached to documents (such as Microsoft Word)
  - Active X controls (programs on web sites with potential access to whole machine)
  - Spyware included with normal products
- Active X controls can have access to the local machine
  - Install software/Launch programs
- Sony Spyware [Sony XCP] (October 2005)
  - About 50 CDs from Sony automatically installed software when you played them on Windows machines
    - » Called XCP (Extended Copy Protection)
    - » Modify operating system to prevent more than 3 copies and to prevent peer-to-peer sharing
  - Side Effects:
    - » Reporting of private information to Sony
    - » Hiding of generic file names of form $sys_xxx; easy for other virus writers to exploit
    - » Hard to remove (crashes machine if not done carefully)
  - Vendors of virus protection software declare it spyware
    - » Computer Associates, Symantec, even Microsoft

Enforcement

- Enforcer checks passwords, ACLs, etc
  - Makes sure the only authorized actions take place
  - Bugs in enforcer⇒things for malicious users to exploit
- In UNIX, superuser can do anything
  - Because of coarse-grained access control, lots of stuff has to run as superuser in order to work
  - If there is a bug in any one of these programs, you lose!
- Paradox
  - Bullet-proof enforcer
    - » Only known way is to make enforcer as small as possible
    - » Easier to make correct, but simple-minded protection model
  - Fancy protection
    - » Tries to adhere to principle of least privilege
    - » Really hard to get right
- Same argument for Java or C++: What do you make private vs public?
  - Hard to make sure that code is usable but only necessary modules are public
  - Pick something in middle? Get bugs and weak protection!

State of the World

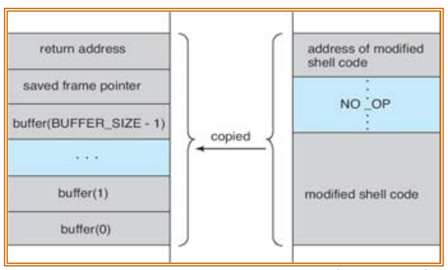
- **State of the World in Security**
  - Authentication: Encryption
    - » But almost no one encrypts or has public key identity
  - Authorization: Access Control
    - » But many systems only provide very coarse-grained access
    - » In UNIX, need to turn off protection to enable sharing
  - Enforcement: Kernel mode
    - » Hard to write a million line program without bugs
    - » Any bug is a potential security loophole!
- **Some types of security problems**
  - Abuse of privilege
    - » If the superuser is evil, we're all in trouble/can't do anything
    - » What if sysop in charge of instructional resources went crazy and deleted everybody's files (and backups)???
  - Imposter: Pretend to be someone else
    - » Example: in unix, can set up an .rhosts file to allow logins from one machine to another without retyping password
    - » Allows "rsh" command to do an operation on a remote node
    - » Result: send rsh request, pretending to be from trusted user→install .rhosts file granting you access

Other Security Problems

- **Virus:**
  - A piece of code that attaches itself to a program or file so it can spread from one computer to another, leaving infections as it travels
  - Most attached to executable files, so don't get activated until the file is actually executed
  - Once caught, can hide in boot tracks, other files, OS
- **Worm:**
  - Similar to a virus, but capable of traveling on its own
  - Takes advantage of file or information transport features
  - Because it can replicate itself, your computer might send out hundreds or thousands of copies of itself
- **Trojan Horse:**
  - Named after huge wooden horse in Greek mythology given as gift to enemy; contained army inside
  - At first glance appears to be useful software but does damage once installed or run on your computer

Security Problems: Buffer-overflow Condition

```
#define BUFFER SIZE 256
int process(int argc,
            char *argv[])
{
  char buffer[BUFFER SIZE];
  if (argc < 2)
       return -1;
  else {
       strcpy(buffer,argv[1]);
       return 0;
  }
}
```

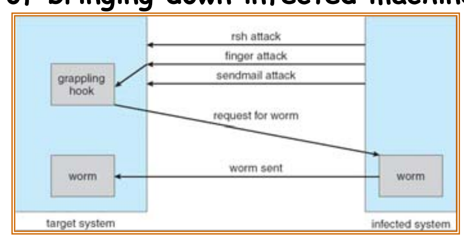


Before attack   After attack

- **Technique exploited by many network attacks**
  - Anytime input comes from network request and is not checked for size
  - Allows execution of code with same privileges as running program – but happens without any action from user!
- **How to prevent?**
  - Don't code this way! (ok, wishful thinking)
  - New mode bits in Intel, Amd, and Sun processors
    - » Put in page table; says "don't execute code in this page"

The Morris Internet Worm

- **Internet worm (Self-reproducing)**
  - Author Robert Morris, a first-year Cornell grad student
  - Launched close of Workday on November 2, 1988
  - Within a few hours of release, it consumed resources to the point of bringing down infected machines



- **Techniques**
  - Exploited UNIX networking features (remote access)
  - Bugs in *finger* (buffer overflow) and *sendmail* programs (debug mode allowed remote login)
  - Dictionary lookup-based password cracking
  - Grappling hook program uploaded main worm program

Some other Attacks

- Trojan Horse Example: Fake Login
  - Construct a program that looks like normal login program
  - Gives "login:" and "password:" prompts
    - » You type information, it sends password to someone, then either logs you in or says "Permission Denied" and exits
  - In Windows, the "ctrl-alt-delete" sequence is supposed to be really hard to change, so you "know" that you are getting official login program
- Salami attack: Slicing things a little at a time
  - Steal or corrupt something a little bit at a time
  - E.g.: What happens to partial pennies from bank interest?
    - » Bank keeps them! Hacker re-programmed system so that partial pennies would go into his account.
    - » Doesn't seem like much, but if you are large bank can be millions of dollars
- Eavesdropping attack
  - Tap into network and see everything typed
  - Catch passwords, etc
  - Lesson: never use unencrypted communication!

Timing Attacks: Tenex Password Checking

- Tenex – early 70's, BBN
  - Most popular system at universities before UNIX
  - Thought to be very secure, gave "red team" all the source code and documentation (want code to be publicly available, as in UNIX)
  - In 48 hours, they figured out how to get every password in the system
- Here's the code for the password check:

```
for (i = 0; i < 8; i++)
  if (userPasswd[i] != realPasswd[i])
    go to error
```

- How many combinations of passwords?
  - 256^8?
  - Wrong!

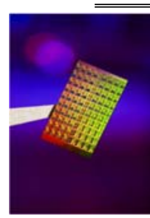
Defeating Password Checking

- Tenex used VM, and it interacts badly with the above code
  - Key idea: force page faults at inopportune times to break passwords quickly
- Arrange 1st char in string to be last char in pg, rest on next pg
  - Then arrange for pg with 1st char to be in memory, and rest to be on disk (e.g., ref lots of other pgs, then ref 1st page)

```
              a|aaaaaa
               |
page in memory| page on disk
```

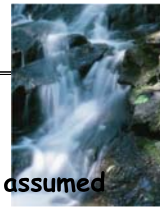
- Time password check to determine if first character is correct!
  - If fast, 1st char is wrong
  - If slow, 1st char is right, pg fault, one of the others wrong
  - So try all first characters, until one is slow
  - Repeat with first two characters in memory, rest on disk
- Only 256 * 8 attempts to crack passwords
  - Fix is easy, don't stop until you look at all the characters

ManyCore Chips: The future is here (for EVERYONE)



- Intel 80-core multicore chip (Feb 2007)
  - 80 simple cores
  - Two floating point engines /core
  - Mesh-like "network-on-a-chip"
  - 100 million transistors
  - 65nm feature size
- "ManyCore" refers to many processors/chip
  - 64? 128? Hard to say exact boundary
- Question: How can ManyCore change our view of OSs?
  - ManyCore is a challenge
    - » Need to be able to take advantage of parallelism
    - » Must utilize many processors somehow
  - ManyCore is an opportunity
    - » Manufacturers are desperate to figure out how to program
    - » Willing to change many things: hardware, software, etc.
  - Can we improve: security, responsiveness, programmability?

PARLab OS Goals: *RAPPidS*

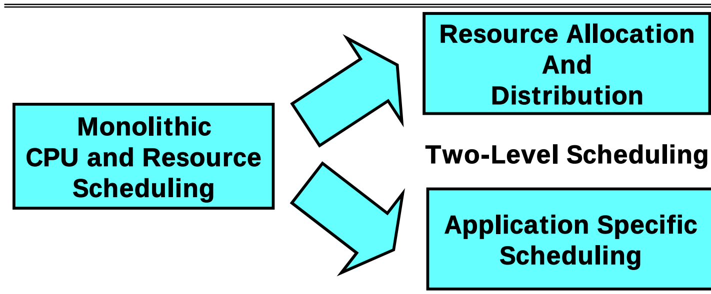


- **R**esponsiveness: Meets real-time guarantees
  - Good user experience with UI expected
  - Illusion of Rapid I/O while still providing guarantees
  - Real-Time applications (speech, music, video) will be assumed
- **A**gility: Can deal with rapidly changing environment
  - Programs not completely assembled until runtime
  - User may request complex mix of services at moment's notice
  - Resources change rapidly (bandwidth, power, etc)
- **P**ower-Efficiency: Efficient power-performance tradeoffs
  - Application-Specific parallel scheduling on Bare Metal partitions
  - Explicitly parallel, power-aware OS service architecture
- **P**ersistence: User experience persists across device failures
  - Fully integrated with persistent storage infrastructures
  - Customizations not be lost on "reboot"
- **S**ecurity and Correctness: Must be hard to compromise
  - Untrusted and/or buggy components handled gracefully
  - Combination of *verification* and *isolation* at many levels
  - Privacy, Integrity, Authenticity of information asserted

The Problem with Current OSs

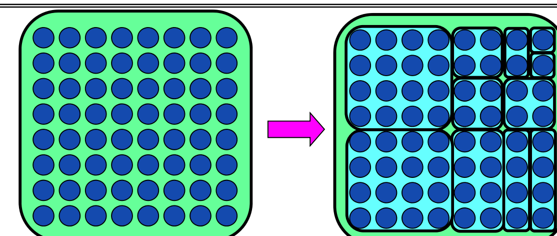
- What is wrong with current Operating Systems?
  - They do not allow expression of application requirements
    - » Minimal Frame Rate, Minimal Memory Bandwidth, Minimal QoS from system Services, Real Time Constraints, …
    - » No clean interfaces for reflecting these requirements
  - They do not provide guarantees that applications can use
    - » They do not provide performance isolation
    - » Resources can be removed or decreased without permission
    - » Maximum response time to events cannot be characterized
  - They do not provide fully custom scheduling
    - » In a parallel programming environment, ideal scheduling can depend crucially on the programming model
  - They do not provide sufficient Security or Correctness
    - » Monolithic Kernels get compromised all the time
    - » Applications cannot express domains of trust within themselves without using a heavyweight process model
- The advent of ManyCore both:
  - Exacerbates the above with greater number of shared resources
  - Provides an opportunity to change the fundamental model

A First Step: Two Level Scheduling



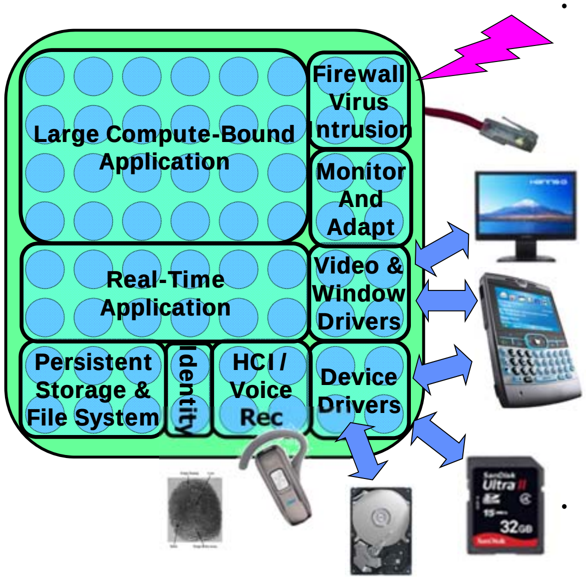
- Split monolithic scheduling into two pieces:
  - Course-Grained Resource Allocation and Distribution
    - » Chunks of resources (CPUs, Memory Bandwidth, QoS to Services) distributed to application (system) components
    - » Option to simply turn off unused resources (Important for Power)
  - Fine-Grained Application-Specific Scheduling
    - » Applications are allowed to utilize their resources in any way they see fit
    - » Other components of the system cannot interfere with their use of resources

Important New Mechanism: Spatial Partitioning



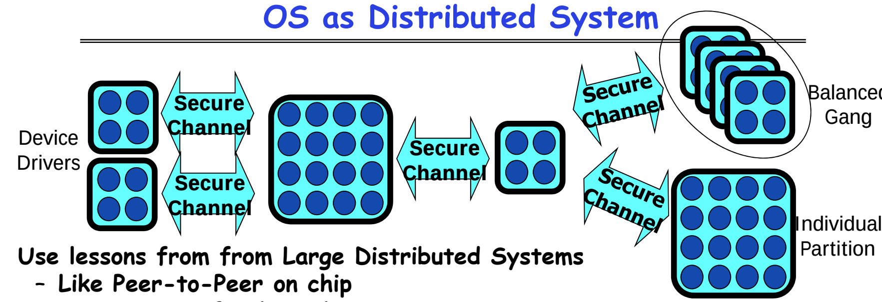
- Spatial Partition: group of processors acting within hardware boundary
  - Boundaries are "hard", communication between partitions controlled
  - Anything goes within partition
- Each Partition receives a *vector* of resources
  - Some number of dedicated processors
  - Some set of dedicated resources (exclusive access)
    - » Complete access to certain hardware devices
    - » Dedicated raw storage partition
  - Some guaranteed fraction of other resources (QoS guarantee):
    - » Memory bandwidth, Network bandwidth
    - » fractional services from other partitions
- Key Idea: Resource Isolation Between Partitions

Tessellation: The Exploded OS



- Normal Components split into pieces
  - Device drivers (Security/Reliability)
  - Network Services (Performance)
    - » TCP/IP stack
    - » Firewall
    - » Virus Checking
    - » Intrusion Detection
  - Persistent Storage (Performance, Security, Reliability)
  - Monitoring services
    - » Performance counters
    - » Introspection
  - Identity/Environment services (Security)
    - » Biometric, GPS, Possession Tracking
- Applications Given Larger Partitions
  - Freedom to use resources arbitrarily

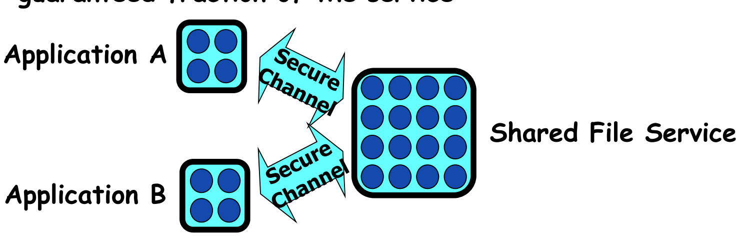
OS as Distributed System



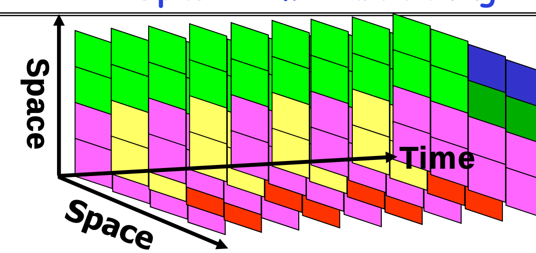
- Use lessons from from Large Distributed Systems
  - Like Peer-to-Peer on chip
  - OS is a set of independent interacting components
  - Shared state across components minimized
- Component-based design:
  - All applications designed with pieces from many sources
  - Requires composition: Performance, Interfaces, Security
- Spatial Partitioning Advantages:
  - Protection of computing resources *not required* within partition
    - » High walls between partitions ⇒ anything goes within partition
    - » "Bare Metal" access to hardware resources
  - Partitions exist simultaneously ⇒ fast communication between domains
    - » Applications split into distrusting partitions w/ controlled communication
    - » Hardware acceleration/tagging for fast secure messaging

It's all about the communication

- We are interested in communication for many reasons:
  - Communication represents a security vulnerability
  - Quality of Service (QoS) boils down message tracking
  - Communication efficiency impacts decomposability
- Shared components complicate resource isolation:
  - Need distributed mechanism for tracking and accounting of resource usage
    - » E.g.: How do we guarantee that each partition gets a guaranteed fraction of the service:



Space-Time Partitioning

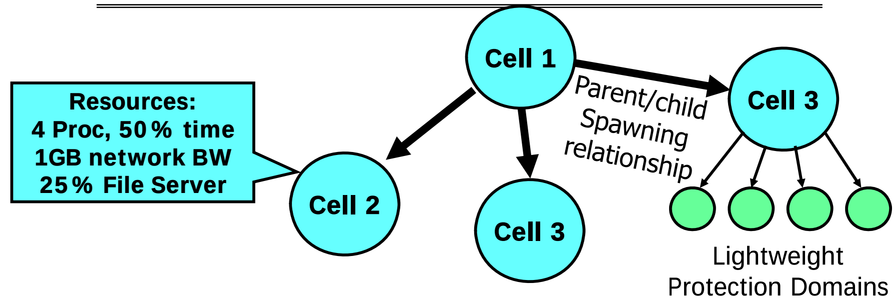


- Spatial Partitioning Varies over Time
  - Partitioning adapts to needs of the system
  - Some partitions persist, others change with time
  - Further, Partititions can be Time Multiplexed
    - » Services (i.e. file system), device drivers, hard realtime partitions
    - » User-level schedulers may time-multiplex threads within partition
- Global Partitioning Goals:
  - Power-performance tradeoffs
  - Setup to achieve QoS and/or Responsiveness guarantees
  - Isolation of real-time partitions for better guarantees
- Monitoring and Adaptation
  - Integration of performance/power/efficiency counters

Another Look: Two-Level Scheduling

- First Level: Gross partitioning of resources
  - Goals: Power Budget, Overall Responsiveness/QoS, Security
  - Partitioning of CPUs, Memory, Interrupts, Devices, other resources
  - Constant for sufficient period of time to:
    - » Amortize cost of global decision making
    - » Allow time for partition-level scheduling to be effective
  - Hard boundaries ⇒ interference-free use of resources
- Second Level: Application-Specific Scheduling
  - Goals: Performance, Real-time Behavior, Responsiveness, Predictability
  - CPU scheduling tuned to specific applications
  - Resources distributed in application-specific fashion
  - External events (I/O, active messages, etc) deferrable as appropriate
- Justifications for two-level scheduling?
  - Global/cross-app decisions made by 1st level
    - » E.g. Save power by focusing I/O handling to smaller # of cores
  - App-scheduler (2nd level) better tuned to application
    - » Lower overhead/better match to app than global scheduler
    - » No global scheduler could handle all applications

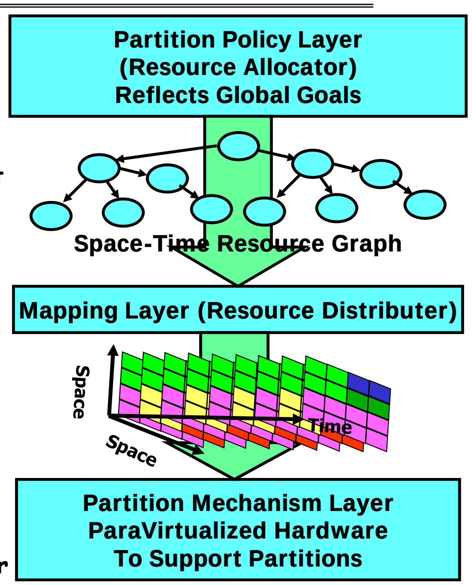
Space-Time Resource Graph



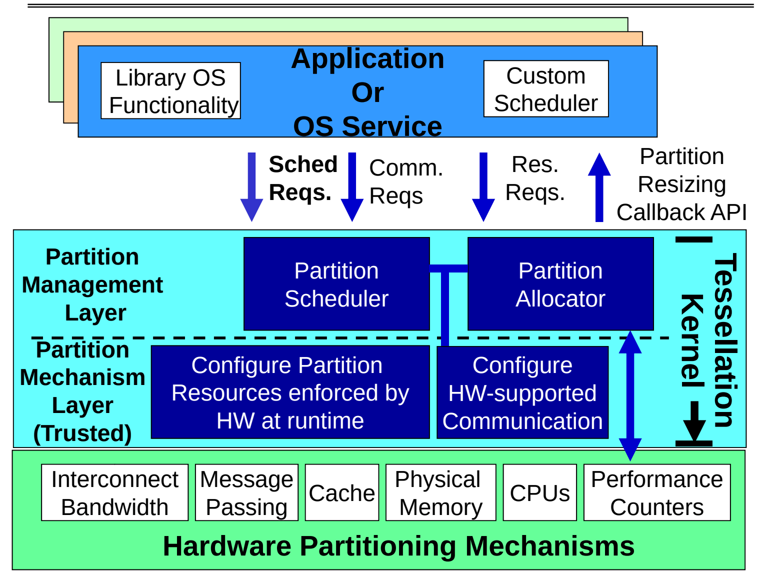
- Space-Time resource graph: the explicit instantiation of resource assignments
  - Directed Arrows Express Parent/Child Spawning Relationship
  - All resources have a Space/Time component
    - » E.g. X Processors/fraction of time, or Y Bytes/Sec
- What does it mean to give resources to a Cell?
  - The Cell has a position in the Space-Time resource graph and
  - The resources are added to the cell's resource label
  - Resources cannot be taken away except via explicit APIs

Implementing the Space-Time Graph

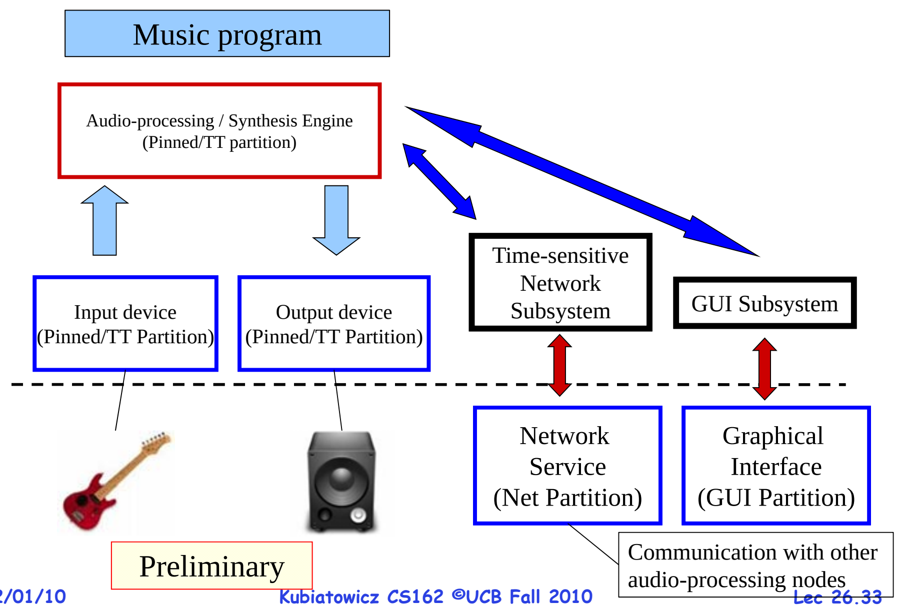
- Partition Policy layer (allocation)
  - Allocates Resources to Cells based on Global policies
  - Produces only implementable space-time resource graphs
  - May deny resources to a cell that requests them (admission control)
- Mapping layer (distribution)
  - Makes no decisions
  - Time-Slices at a course granularity
  - performs bin-packing like to implement space-time graph
  - In limit of *many* processors, no time multiplexing processors, merely distributing resources
- Partition Mechanism Layer
  - Implements hardware partitions and secure channels
  - Device Dependent: Makes use of more or less hardware support for QoS and Partitions



Tessellation Architecture



Example of Music Application



Conclusion

- Distributed identity
  - Use cryptography (Public Key, Signed by PKI)
- Distributed storage example
  - Revocation: How to remove permissions from someone?
  - Integrity: How to know whether data is valid
  - Freshness: How to know whether data is recent
- Buffer-Overrun Attack: exploit bug to execute code
- Space-Time Partitioning: grouping processors & resources behind hardware boundary
  - Focus on Quality of Service
  - Two-level scheduling
    1) Global Distribution of resources
    2) Application-Specific scheduling of resources
  - Bare Metal Execution within partition
  - Composable performance, security, QoS
- Tessellation Paper:
  - Off my "publications" page (near top): http://www.cs.berkeley.edu/~kubitron/papers

Good Bye!

- Optional Lecture on Monday
- Let's thank the TAs!
- Good Bye!
  You have been a great class!