Virtual synchrony

Ken Birman

Virtual Synchrony

- Goal: Simplifies distributed systems development by introducing emulating a simplified world – a *synchronous* one
- Features of the virtual synchrony model
  - Process groups with state transfer, automated fault detection and membership reporting
  - Ordered reliable multicast, in several flavors
  - Fault-tolerance, replication tools layered on top
  - Extremely good performance

Process groups

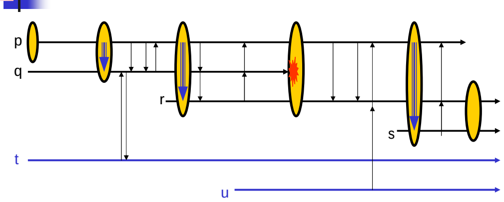
- Offered as a new and fundamental programming abstraction
  - Just a set of application processes that cooperate for some purpose
  - Could replicate data, coordinate handling of incoming requests or events, perform parallel tasks, or have a shared perspective on some sort of "fact" about the system
- Can create many of them*

** Within limits... Many systems only had limited scalability*

Why "virtual" synchrony?

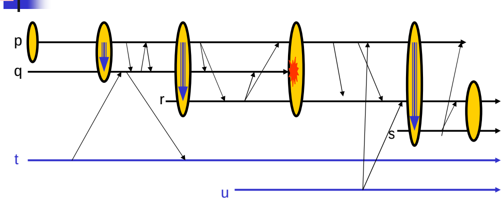
- What would a synchronous execution look like?
- In what ways is a "virtual" synchrony execution not the same thing?

A synchronous execution



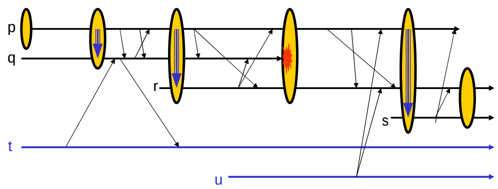
- With *true* synchrony executions run in genuine lock-step.

Virtual Synchrony at a glance



With *virtual* synchrony executions only look "lock step" to the application

Virtual Synchrony at a glance



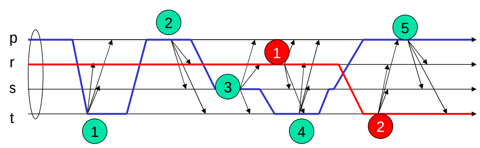
We use the weakest (least ordered, hence fastest) form of communication possible

Chances to "weaken" ordering

- Suppose that any conflicting updates are synchronized using some form of locking
  - Multicast sender will have mutual exclusion
  - Hence simply because we used locks, cbcast delivers conflicting updates in order they were performed!
- If our system ever *does* see concurrent multicasts… they must not have conflicted. So it won't matter if cbcast delivers them in different orders at different recipients!

Causally ordered updates

- Each thread corresponds to a different lock



- In effect: red "events" never conflict with green ones!

In general?

- Replace "safe" (dynamic uniformity) with a standard multicast when possible
- Replace abcast with cbcast
- Replace cbcast with fbcast
- Unless replies are needed, don't wait for replies to a multicast

Why "virtual" synchrony?

- The user writes code as it will experience a purely synchronous execution
  - Simplifies the developer's task – very few cases to worry about, and all group members see the same thing at the same "time"
- But the actual execution is rather concurrent and asynchronous
  - Maximizes performance
  - Reduces risk that lock-step execution will trigger correlated failures

Why groups?

- Other concurrent work, such as Lamport's state machines, treat the entire program as a deterministic entity and replicate it
- But a group replicates state at the "abstract data type" level
  - Each group can correspond to one object
- This is a good fit with modern styles of application development

Correlated failures

- Perhaps surprisingly, experiments showed that virtual synchrony makes these less likely!
  - Recall that many programs are buggy
  - Often these are Heisenbugs (order sensitive)
- With lock-step execution each group member sees group events in identical order
  - So all die in unison
- With virtual synchrony orders differ
  - So an order-sensitive bug might only kill one group member!

Programming with groups

- Many systems just have one group
  - E.g. replicated bank servers
  - Cluster mimics one highly reliable server
- But we can also use groups at finer granularity
  - E.g. to replicate a shared data structure
  - Now one process might belong to many groups
- A further reason that different processes might see different inputs and event orders

Embedding groups into “tools”

- We can design a groups API:
  - pg_join(), pg_leave(), cbcast()...
- But we can also use groups to build other higher level mechanisms
  - Distributed algorithms, like snapshot
  - Fault-tolerant request execution
  - Publish-subscribe

Distributed algorithms

- Processes that might participate join an appropriate group
- Now the group view gives a simple leader election rule
  - Everyone sees the same members, in the same order, ranked by when they joined
  - Leader can be, e.g., the “oldest” process

Distributed algorithms

- A group can easily solve consensus
  - Leader multicasts: “what’s your input”?
  - All reply: “Mine is 0. Mine is 1”
  - Initiator picks the most common value and multicasts that: the “decision value”
  - If the leader fails, the new leader just restarts the algorithm
- Puzzle: Does FLP apply here?

Distributed algorithms

- A group can easily do consistent snapshot algorithm
  - Either use cbcast throughout system, or build the algorithm over gbcast
  - Two phases:
    - Start snapshot: a first cbcast
    - Finished: a second cbcast, collect process states and channel logs

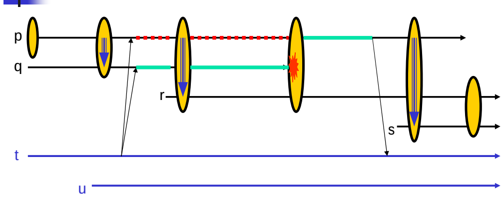
Distributed algorithms: Summary

- Leader election
- Consensus and other forms of agreement like voting
- Snapshots, hence deadlock detection, auditing, load balancing

More tools: fault-tolerance

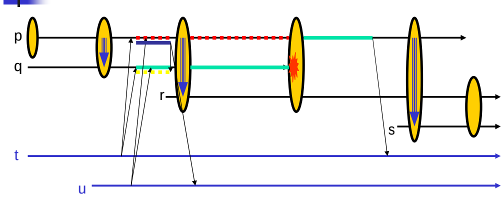
- Suppose that we want to offer clients "fault-tolerant request execution"
  - We can replace a traditional service with a group of members
  - Each request is assigned to a primary (ideally, spread the work around) and a backup
    - Primary sends a "cc" of the response to the request to the backup
  - Backup keeps a copy of the request and steps in only if the primary crashes before replying
- Sometimes called "coordinator/cohort" just to distinguish from "primary/backup"

Coordinator-cohort



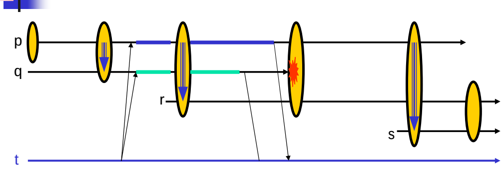
Q assigned as coordinator for t's request.... But p takes over if q fails

Coordinator-cohort



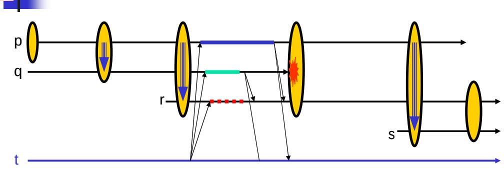
P picked to perform u's request. Q stands by until it sees request completion message

Parallel processing



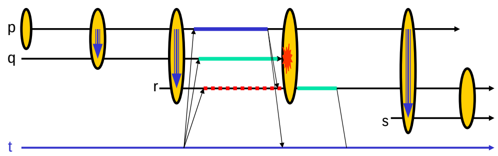
P and Q split a task. P performs part 1 of 2; Q performs part 2 of 2. Such as searching a large database... they agree on the initial state in which the request was received

Parallel processing



In this example, r is the cohort and both p and q function as coordinators. If *either* fails, r can step in and take over its role....

Parallel processing



.... As it did in this case, when q fails

Publish / Subscribe

- Goal is to support a simple API:
  - Publish("topic", message)
  - Subscribe("topic", event_hander)
- We can just create a group for each topic
  - Publish multicasts to the group
  - Subscribers are the members

Scalability warnings!

- Many existing group communication systems don't scale incredibly well
  - E.g. JGroups, Isis, Horus, Ensemble, Spread
  - Group sizes limited to perhaps 50-75 members
  - And individual processes limited to joining perhaps 50-75 groups (lightweight groups an exception)
- Overheads soar as these sizes increase
  - Each group runs protocols oblivious of the others, and this creates huge inefficiency

Publish / Subscribe issue?

- We could have thousands of topics!
  - Too many to directly map topics to groups
- Instead map topics to a *smaller set* of groups.
  - SPREAD system calls these "lightweight" groups (idea traces to work done by Glade on Isis)
  - Mapping will result in inaccuracies... Filter incoming messages to discard any not actually destined to the receiver process
- Cornell's new QuickSilver system instead directly supports immense numbers of groups

Other "toolkit" ideas

- We could embed group communication into a framework in a "transparent" way
  - Example: CORBA fault-tolerance specification does lock-step replication of deterministic components
  - The client simply can't see failures
    - But the determinism assumption is painful, and users have been unenthusiastic
    - And exposed to correlated crashes

Other similar ideas

- There was some work on embedding groups into programming languages
  - But many applications want to use them to link programs coded in different languages and systems
  - Hence an interesting curiosity but just a curiosity
- Quicksilver: Transparently embeds groups into Windows

Existing toolkits: challenges

- Tensions between threading and ordering
  - We need concurrency (threads) for perf.
  - Yet we need to preserve the order in which "events" are delivered
- This poses a difficult balance for the developers

Features of major virtual synchrony platforms

- Isis: First and no longer widely used
  - But was the most successful; has major roles in NYSE, Swiss Exchange, French Air Traffic Control system (two major subsystems of it), US AEGIS Naval warship
  - Also was first to offer a publish-subscribe interface that mapped topics to groups

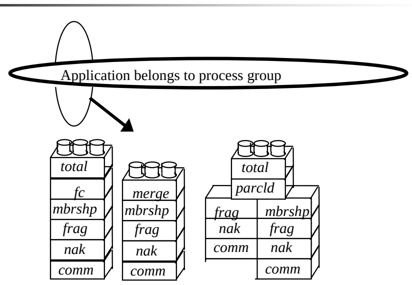
Features of major virtual synchrony platforms

- Totem and Transis
  - Sibling projects, shortly after Isis
  - Totem (UCSB) went on to become Eternal and was the basis of the CORBA fault-tolerance standard
  - Transis (Hebrew University) became a specialist in tolerating partitioning failures, then explored link between vsync and FLP

Features of major virtual synchrony platforms

- Horus, JGroups and Ensemble
  - All were developed at Cornell: successors to Isis
  - These focus on flexible protocol stack linked directly into application address space
    - A stack is a pile of micro-protocols
    - Can assemble an optimized solution fitted to specific needs of the application by plugging together "properties this application requires", lego-style
    - The system is optimized to reduce overheads of this compositional style of protocol stack
  - JGroups is *very* popular.
  - Ensemble is somewhat popular and supported by a user community. Horus works well but is not widely used.

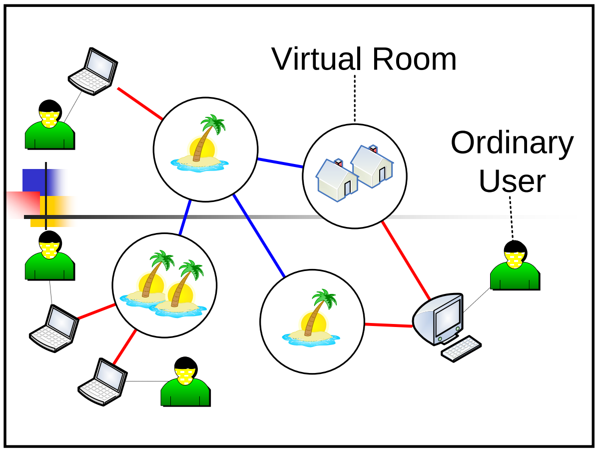
Horus/JGroups/Ensemble protocol stacks

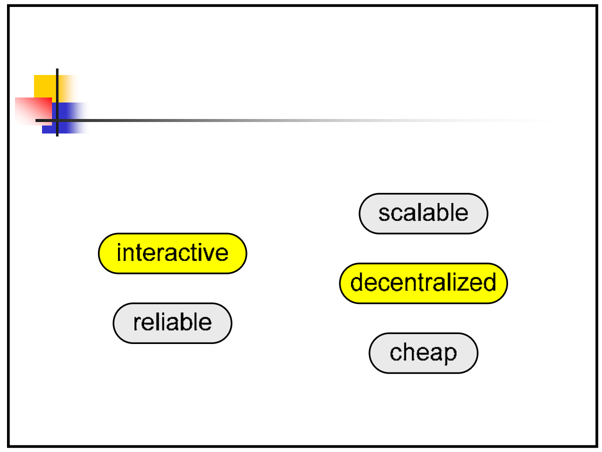


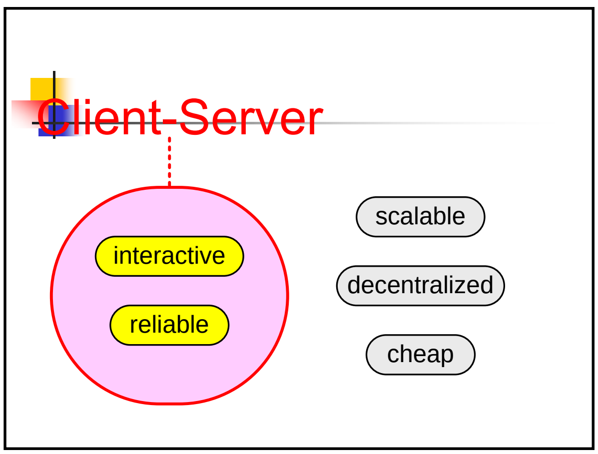
QuickSilver Scalable Multicast

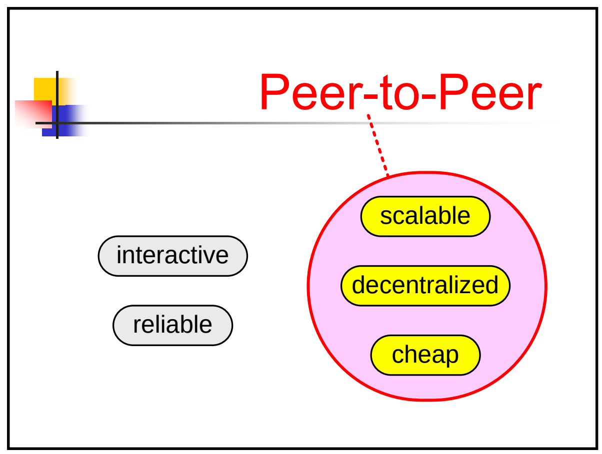
Thinking beyond Web 2.0

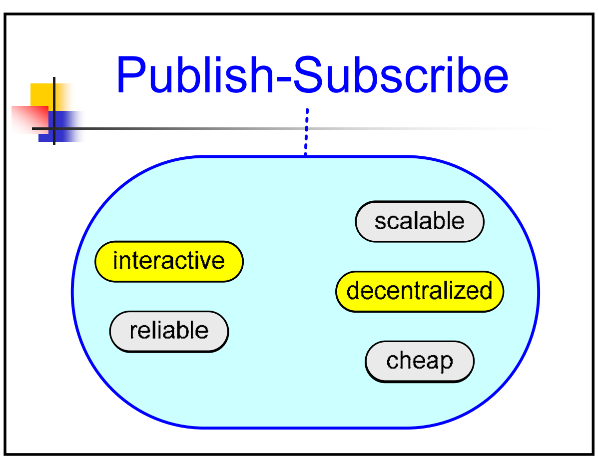
Krzysztof Ostrowski, Ken Birman
Cornell University
{krzys,ken}@cs.cornell.edu

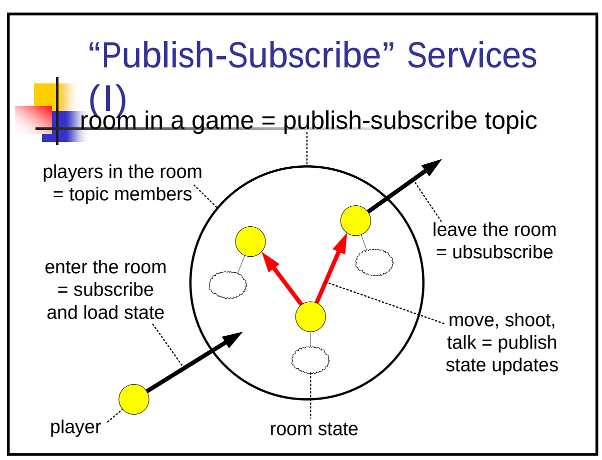

Virtual Room
Ordinary User


interactive
reliable
scalable
decentralized
cheap

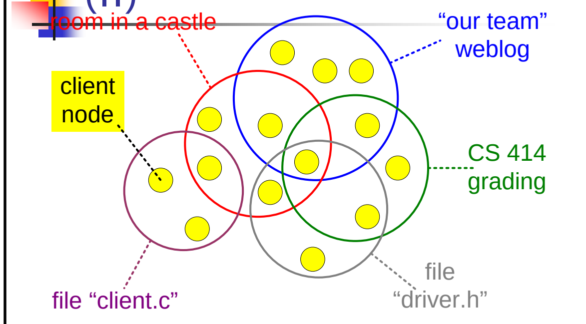

Client-Server
interactive
reliable
scalable
decentralized
cheap


Peer-to-Peer
interactive
reliable
scalable
decentralized
cheap


Publish-Subscribe
interactive
reliable
scalable
decentralized
cheap


“Publish-Subscribe” Services (I)
room in a game = publish-subscribe topic
players in the room = topic members
enter the room = subscribe and load state
leave the room = ubsubscribe
move, shoot, talk = publish state updates
player
room state

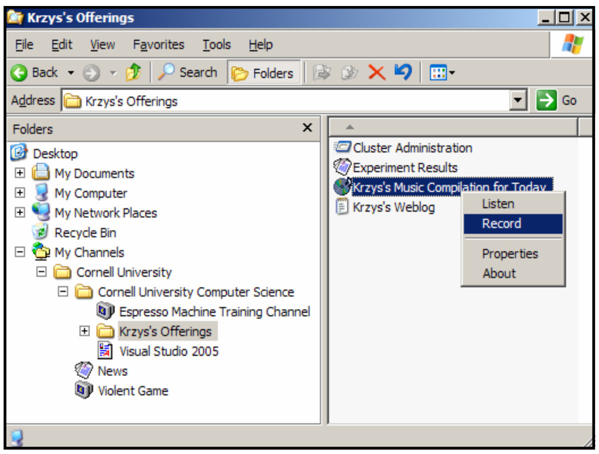
“Publish-Subscribe” Services (II)



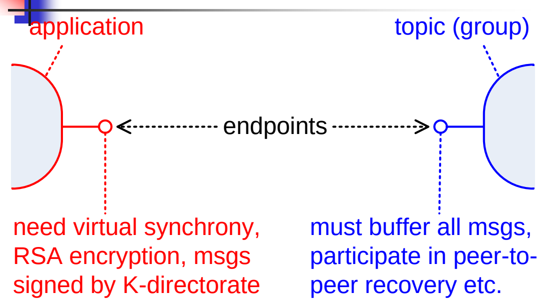
New Style of Programming

Topics = Objects

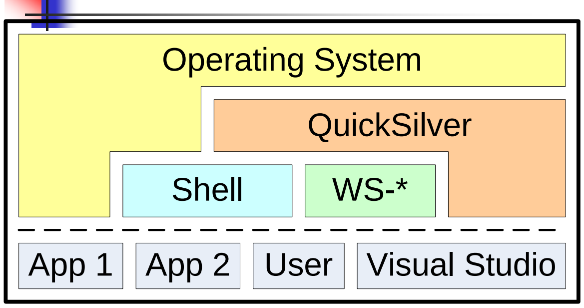
```
Topic x = Internet.Enter(“Game X”);
Topic y = x.Enter(“Room X”);
y.OnShoot +=
  new EventHandler(this.TurnAround);
while (true)
  y.Shoot(new Vector(1,0,0));
```

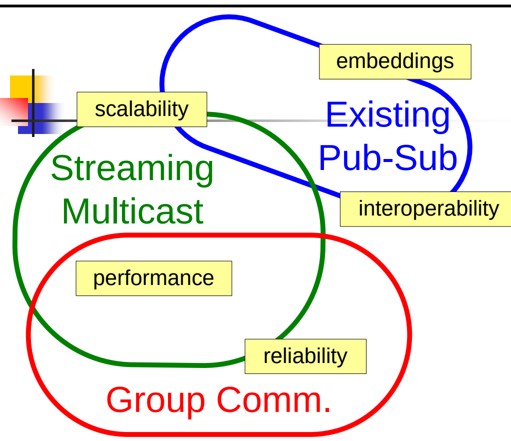


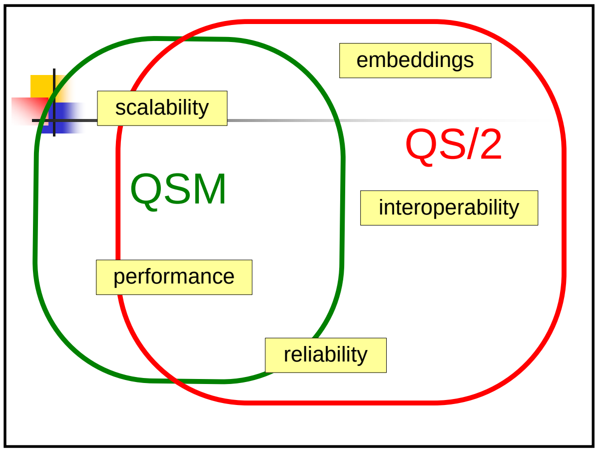
Typed Publish-Subscribe



Where does QuickSilver belong?

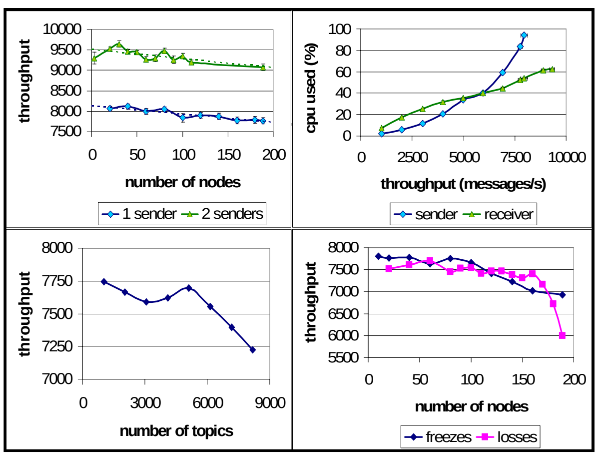






QuickSilver Scalable Multicast

- Simple ACK-based reliability property
- Managed code (.NET, 95%C#, 5%MC++)
- Entire QuickSilver platform: ~250 KLOC
- Throughputs close to network speeds
- Scalable in multiple dimensions
- Tested with up to ~200 nodes, 8K groups
- Robust against a range of perturbances



Summary?

- Role of a toolkit is to package commonly used, popular functionality into simple API and programming model
- Group communication systems have been more popular when offered in toolkits
  - If groups are embedded into programming languages, we limit interoperability
  - If groups are used to transparently replicate deterministic objects, we're too inflexible
- Many modern systems let you match the protocol to your application's requirements