

# Survivable Information Access

Johns Hopkins  
Yair Amir, Claudiu Danilov,  
Jonathan Kirsch, John Lane

Purdue  
Chi-Bun Chan, Cristina Nita-  
Rotaru, Josh Olsen, David Zage

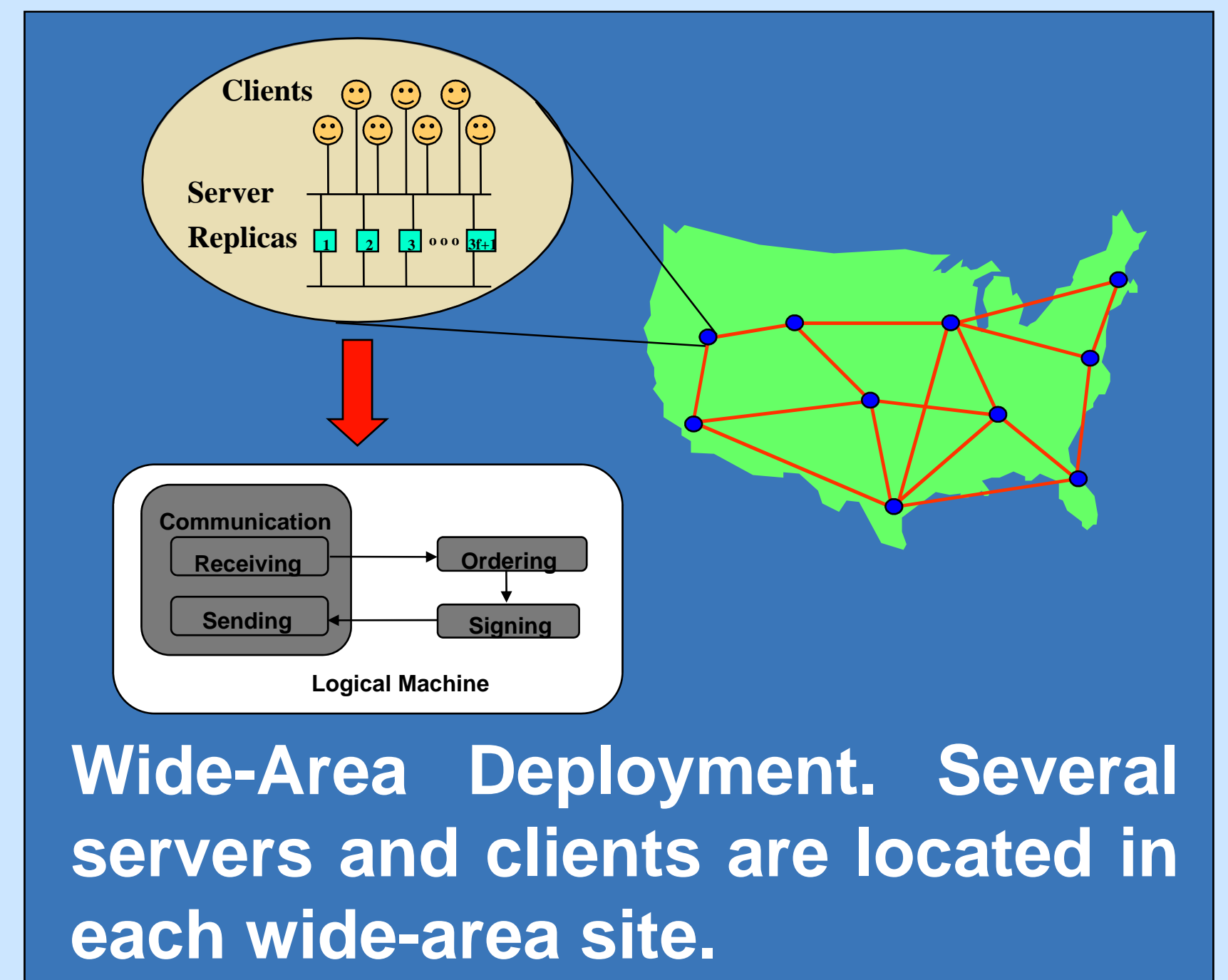
Telcordia Technologies  
Brian Coan

## Problem

Byzantine replication can be used to build systems that survive insider attacks mounted by compromised servers. Previous solutions performed well in local-area networks. Our architecture, Steward, was the first to scale Byzantine fault-tolerant replication to wide-area networks, where servers are located in many sites distributed across the Internet. Steward's architecture reduces communication costs, enabling it to achieve performance an order of magnitude above the previous state of the art. It successfully met safety (data consistency) and liveness (eventual progress) guarantees even during a white-box red-team experiment where a knowledgeable attacker was given full control of some of the servers. Steward's performance came at a cost: inflexibility and complexity. The protocols used within the local-area sites and the protocol used on the wide area were tightly coupled, making it impossible to customize the fault-tolerance approach used within and among the sites. To address this problem, we developed a new composable Byzantine wide area replication architecture that provides a clean separation between the protocols used within and among the sites.

## Scalable Byzantine Replication

- The previous state of the art (Castro-Liskov's BFT) was designed for use in local-area networks, and requires costly all-to-all message exchanges.
- Our systems are designed for large deployments where servers and clients are located in several local-area sites distributed across a wide-area network.
- We leverage hierarchy to (1) reduce the number of messages sent on the wide area and (2) allow queries to be answered locally.



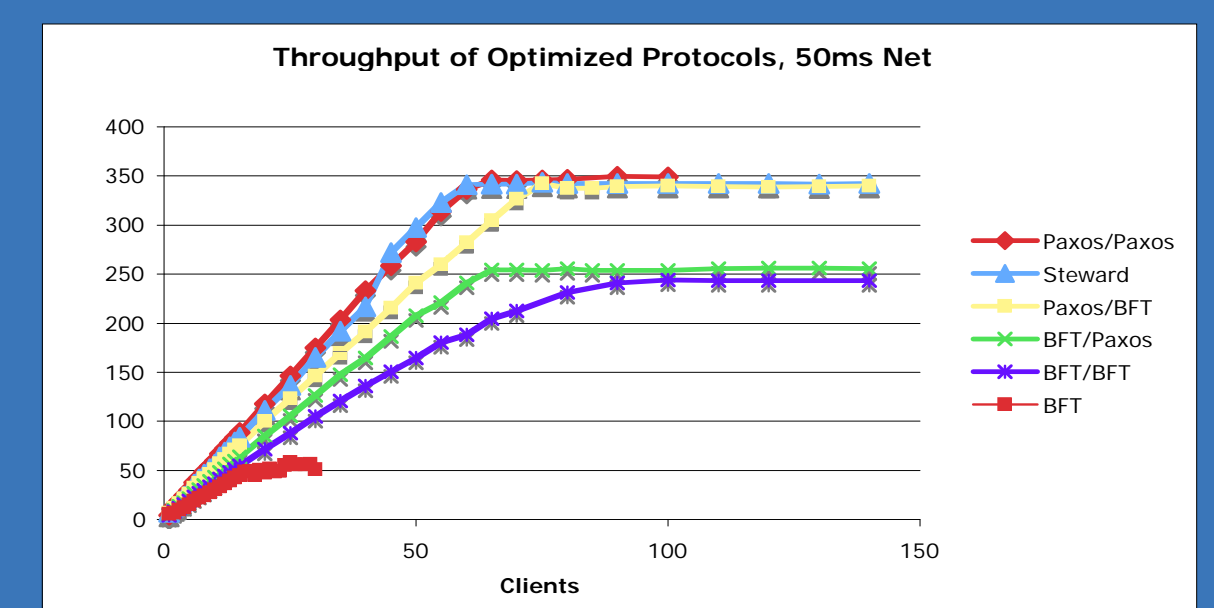
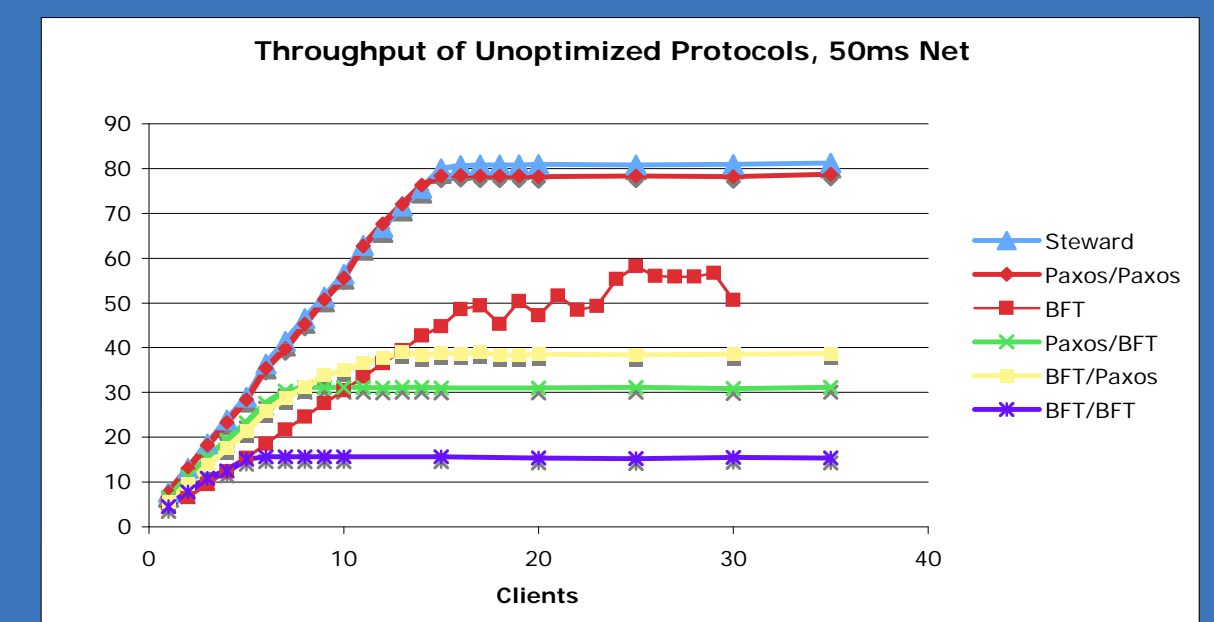
## Comparison to State of the Art

### Castro-Liskov's BFT

- Flat architecture requires messages to be sent between all servers in the system.
- Survives compromises of  $f$  malicious servers out of a total of  $3f+1$  servers.
- Wide-area message complexity of  $O(N^2)$  where  $N$  is the number of servers.
- Queries require that messages be sent on the wide area.
- Requires 3 rounds of wide-area communication.

### Our Approach

- Hierarchical architecture tailored for efficient use of wide-area bandwidth.
- Survives compromise of up to (but not including) one third of servers in each site. Composable Architecture survives complete site compromises.
- Wide-area message complexity of  $O(S^2)$  where  $S$  is the number of sites.
- Queries can be answered locally, improving both performance and availability.
- Either 2 or 3 wide-area rounds, depending on wide-area fault tolerance.



## Steward and the Composable Architecture -- Protocol Highlights

- **Physical machines in each site act as a logical entity that plays the role of a single participant in a wide-area replication protocol.**
  - **Steward:** Servers within a site run Byzantine local-area protocols to mask the effects of malicious servers. A benign fault-tolerant protocol (resilient to servers crashes and network partitions) runs among the sites.
  - **Composable Architecture:** Servers in each site implement a **logical machine** by running a local state machine replication protocol, and a wide-area replication protocol runs among the logical machines. Affords **free substitution of fault tolerance used within and among sites.**
- **Wide-area protocol messages are threshold signed**
  - Attests that at least  $f+1$  servers in a site agreed on the content.
  - Malicious servers cannot forge messages from a site.
- **Composable Architecture uses optimizations to improve performance**
  - **Byzantine link protocol (BLink)** efficiently sends messages between logical machines.
  - Aggregation and Merkle hash trees amortize the cost of public key cryptography.

