If Iterative Diffusion Is The Answer, What Was The Question?

David Guzman, Dirk Trossen, Joerg Ott

ACM SIGCOMM Workshop on the Future of Internet Routing and Addressing

September, 2023
Framing our Discussions Today

• Distributed consensus (DC) is key for (distributed) computation
• Achieving consensus is inherently a multipoint operation
• We discuss today
  • Why multicast has not been used?
  • What is the current answer for realizing multipoint operations in DCSs?
    • Empirical data
    • Model
• We present a Gedankenexperiment on possible multicast gains
• We then discuss what multicast issues remain to be addressed
The Need for Distributed Consensus

• In a distributed system it is key to agree on a (distributed) state change to continue its evolution (computation)

• Hence, communicating state changes among peer participants is paramount to achieving consensus

• In fact, it is required to propagate information to at least half of the participant peers, the so-called majority rule [VonNewman1956]
Framing the DCS Problem

• A DCS aims at diffusing/disseminating state information, from a source to at least half of the system, 1-to-$S/2$

• Three performance aspects are crucial
  • **Latency** to communicate state changes to achieve the majority
  • The **finality** to diffuse (to at least $S/2$ peers) is critical to conclude consensus
  • At scale, the communication **costs** are important for deployment and non-negligible [Guzman2022]

• **Key Question:** How to diffuse information at (Internet) scale?
  • It sounds like a perfect application for IP multicast due to the opportunities in **efficiency** and **scale** of operation!
  • ...or not?
Why Did We End Up NOT Using Multicast

• **Intra-domain deployments** have been made, driven, e.g., by (local) IPTV use cases
  • ...and link-local multicast is well used in many use cases
  • ...and wide-area tunnels may even interconnect multicast islands [Farinacci2018]

• However, there is a lack of **inter-domain deployment** of IP multicast
  • Economic incentives [Diot2000]: receivers and access network providers do not benefit from the peering providers business models
  • Technological issues [Crowcroft2022]:
    • **Security** of the global Ethernet, i.e., (distributed) denial of service
    • **Scale** in (inter-domain) routing algorithms
    • Content layer (routing) **alignment** with network layer (ISP Policy)
Understanding the Current Answer
Iterative Diffusion in P2P Network

- We consider a system of $S$ total peers, i.e., required to diffuse information to at least $m > S/2$ peers
- A diffusion is a recursive process where $t_d(N)$ is the time for each diffusion
- $N$ is the size of the diffusion set at each iteration
- Each diffusion step is unique with a given probability $u(t)$, i.e., $u(t) = 1$ means that each diffusion step would be unique
- We build a model based on $N, t_d(N), u(t)$ to determine $t_m$ and $t_S$ with confidence intervals
- The diffusion evolution over timeslots $t$ is:

$$m_{tc} \geq (u(t)N)^{\frac{t}{td(N)}} \quad (1)$$

Two routes of investigation: bounding the diffusion latency and the uniqueness
Understanding the Current Answer
Bounding The Diffusion Latency

- Three key steps in ensuring diffusion communication [Guzman2022]: ensuring peer availability, reachability, and capability support

- Empirical measurements yield a Long-tailed PDF due to echo timeouts, TCP connection establishment, and capability timeouts

- We preserve the main behavior in our approximation, namely replenishing the pool of $N$ peers
  - Peer availability & reachability, albeit not rare ones such as capability mismatch

- Note that with some probability albeit small, there is a chance for long diffusion times $t_d(N)$, and thus also for high variance convergence time

$$t_d(N) \sim \ln(\text{Normal}(\mu_d, \sigma_d^2))$$
Understanding the Current Answer
Bounding The **Uniqueness** of Diffusion Peers

- Uniqueness is directly impacted by the selection of the set of peers for each diffusion iteration
- Initially, we infer the probability distribution for $u(t)$ from local empirical data based on a set of peer IDs*
  - We test the selection by spreading a locally generated ID in the system, and taking samples of the incoming and outgoing relations
- Evaluate the selection sequence of set of peers by finding its intersection, and defining the proportion of intended $N$ peers as: $(N - \#repeated\_peers)/N$
- We show that the mean behavior for peer selection can be approximated by a normal distribution, as
  $$u(t) \sim N(\mu_u, \sigma_u^2)$$

*Peer IDs are determined by the MSBs of the digest over a public key generated by local random generators.
Bringing it Back Together: Our **Results**

- Model with parameters
  - $N = 50$ [peers] (standard configuration in Ethereum)
  - $S = 72k$ [peers] (active peers in Ethereum)
  - $m = 36k$ [peers] (peers needed for majority rule)
  - $u(t) = 0.97$

- For mean $t_d(N) \approx 3.7s$, convergence takes place in $t_c \approx 10s$, which matches studies reported in [Zhang2021]

- Rare events cause large variance of $t_c$, causing up to $t_c \approx 26.5s$ convergence latency.
What if Multicast was the Answer?

Gedankenexperiment

- We take a convergence channel $G$ and a single packet transaction* to describe a $(\hat{S}, G)$ multicast to $S/2 + s$ [peers].
- Delivery trees towards $S/2$ (solution specific).
- Group membership to $S$, aggregation aiding scalability.
- **Convergence** happens in a single forward operation 1-to-$(S/2 + s)$  
  - Latency then can be considered to be bound by Internet latency
  - Avoiding random iterative steps based on local (expensive) maintained pools
- **Takeaway**: convergence time up to 10 times or more, lower  
  - Tremendous throughput of, e.g., cryptocurrency, transactions in the overall system
- **Also**: bounded finality of the diffusion operation

* ~256 B in ETH, ~247 B in BTC
Addressing The Multicast Issues

• **Economic incentives**
  • Use of application- instead of network layer multicast may avoid the economic disincentives
  • *new* use cases together with the constant growth of energy consumption/efficiency awareness open the possibility for new commercial uses

• **Technological issues**
  • Distributed Denial of Service (DDoS): *group management* for permissionless peers
  • Billing for data: [Henderson2000]
    • Need to aggregate topological information for downstream accountability
    • Group membership aids in aggregating topological aspects
  • Investigate more efficient shim layer solutions, such as BIER, SD-WAN and similar

• **Even if we solve these issues, is a (global) multicast condemned to fail?**
  • May want to look into *hybrid/app-layer* multicast instead of IP multicast
  • ...and learn from previous lessons in global IP multicast rollouts
Takeaway

• DCS are growing in scale and **importance**
• Despite its **multipoint** nature, DCSs are implemented through **unicast**
  • Leading to high **convergence latency** and **variance** as well as **costs**
• We provided insights into
  • Why **multicast** has not been used for DCSs at scale
  • A **model** that allows for bounding the convergence latency in today’s system, and
  • A **Gedankenexperiment** that framed the possible gain if we did use multicast instead
• However, we will still need to think well about the pitfalls of the past
  • inter-domain deployments, security, economic incentives and pricing
• **Future work:** deepen the Gedankenexperiment with further empirical data and theoretical insights
Thank you.

Bring digital to every person, home and organization for a fully connected, intelligent world.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.