

Ably's Four Pillars of Dependability

Ably is mathematically modelled and architected to overcome limitations of message ordering and delivery without sacrificing latencies, fault tolerance, or service availability.

We can transparently define and measure these operating boundaries of Ably, which we organize under four pillars. Designing around these pillars allows us to remove much of the uncertainty and application complexity usually present when developing apps for realtime.

Developers building on Ably can therefore efficiently design, quickly ship, and seamlessly scale critical realtime functionality ready for global-scale production without huge upfront engineering headaches or runaway ongoing infrastructure costs.

| | |
|---|--|
| Performance of messages We focus on predictability of latencies to provide certainty in uncertain operating conditions. <i><65ms roundtrip latency in 99% percentile Unlimited channel throughput</i> | Integrity of data Guarantees for ordering and delivery to overcome limitations of pub/sub & simplify app architecture. <i>Message Ordering and Delivery 100% guaranteed from publisher to subscribers</i> |
| Reliability of infrastructure Fault tolerant at regional and global level so we can survive multiple failures without any outages. <i>99.999999% message survivability for instance and datacenter failure</i> | Availability of service A transparent, mathematically grounded design for extreme scale, elasticity, and service uptime. <i>Capacity margin for instant surge 99.999% uptime SLA</i> |

Performance of messages

“Performance is the roundtrip, end-to-end latency and bandwidth requirement of sending data.” Dr Paddy Byers, CTO of Ably.

Performance isn't simply about minimizing latency and bandwidth requirements. It's also about minimizing the variance in them and providing predictability to developers. If you know Ably's median global latencies and bandwidth will always be within specific operating boundaries it provides a level of certainty in uncertain operating conditions. You can design, build, and scale features around this certainty, confident they'll perform as expected under various conditions.

- **Round trip latency from any of our 200 PoPs globally that receive at least 1% of our global traffic: target < 65ms with 99% percentile**

This represents the transit latency individual clients experience. Measured at the Point of Presence (PoP) boundary within the Ably access network, which will be closer than a datacenter. Limited to PoPs that receive at least 1% of global traffic.

- **Channel throughput: 200 messages per second, 13MiB per second**

Ably provides unlimited throughput capacity at a system-level. We achieve this with channels (sharding), constraining message and bandwidth throughput per second at a channel (shard) level in order to provide predictable performance. You can activate an unlimited number of channels for unlimited throughput.

- **Channel resource allocation latency: target <200ms, 99% percentile**

Ably is a stateful system so there is a latency 'cost' when activating new channels. This latency has no material impact on the performance for subscribers and is different from message round-trip latency (< 65ms), which is the latency subscribers actually experience. It's possible to activate channels ahead of time to bypass this initial resource allocation latency and increase predictability of latency for clients.

- **Channel churn rate: limitless (constrained only by quota)**

This is the rate at which you can allocate and deallocate channels. This is effectively limitless: you can in theory activate one million channels per second.



Integrity of data

“Integrity comes from the guarantees we provide around realtime messages sent using the Ably service.” Dr Paddy Byers, CTO of Ably.

When apps rely on a sequence of messages that mutually depend upon one another, like chat, Ably maintains the end-to-end integrity of them. This simplifies app architecture: there’s no need to handle missed, unordered, or duplicate messages. This frees you from design limitations so you can focus on solving the challenges that really matter, not the frustrating realtime edge cases you’re otherwise forced to think about and develop around.

- **Guaranteed Message Ordering from any single realtime or non-realtime publisher to all subscribers**

Simplify app architecture and development as Ably ensures message ordering, so you don’t need to handle unordered messages. Customers like HubSpot, Vitac, Genius Sports, and 17Media rely on Ably for ordering so they can simplify their engineering.

- **100% Guaranteed Message Delivery and Onwards Processing**

Ably’s design and protocol ensures that once an ACK is received by the publishing client, all subscribers on that channel are guaranteed to receive the message.

- **Idempotent publish operations are guaranteed within two minutes**

We guarantee messages will be published only once as we discard those delivered multiple times. This provides flexibility around how you design your app as you don’t need to account for duplication. Limited to two minutes as we are a realtime service.

- **Exactly-once semantics with the Ably protocol**

Ably’s exactly-once semantics mean you can simplify your app so it doesn’t need to account for multiple message deliveries, as is the case with at least once or at most once delivery. This is dependent on the reliability of both Ably and the consumer.

- **Preserve connection guarantee across disconnection for two minutes**

Ably ensures connection state is maintained so abrupt disconnections or intermittent connections are resumed automatically by the SDKs, and message stream continuity ensured. Messages published when disconnected are delivered upon reconnection.



Reliability of infrastructure

“Reliability is the ability to continue operating in spite of something going wrong.” Dr Paddy Byers, CTO of Ably.

Ably’s platform is fault tolerant at global and regional levels. We’ve designed around statistical risks of failure, ensuring sufficient redundancy in our infrastructure to ensure continuity of service even in the face of multiple infrastructure failures. Companies like Split.io build on Ably because they know our system is designed in such a way that even if we are facing issues, the statistical risk of issues affecting their end-users is immaterial.

- **[Regional] Message survivability of as a result of instance failures**

We immediately begin migrating messages to two Availability Zones (AZs) so we can replicate them. We design so instance failure doesn’t affect this. We calculate instance failure on the fact that any two instances failing within a five minute window of one another is 0.0000007%. Any instance failure, we migrate to a healthy instance within eight seconds. We offer **99.999999% (8x9s) message survivability**.

- **[Regional] Message survivability as a result of datacenter failure**

If there’s a problem causing issues within an AZ, for example a networking issue, we won’t be able to redistribute load within a datacenter. In this case, we fall back to datacenters in other AZs. We can survive two AZs going down simultaneously without bringing more AZs online. Ably is designed around AZs with 99.99% SLAs, which statistically means we can provide **99.999999% (8x9s) message survivability**.

- **[Global] Persisted data survivability as a result of regional failures**

This measures the reliability of our globally-available long-term storage. Once messages are persisted, we provide **99.99999999% (10x9s) survivability**. You can continue to access data even if one or more regions globally might be down.

- **[Global] Edge network failure resolution by client SDKs within 30s**

Our SDKs can detect and resolve faults by finding a healthy datacenter within 30s.

- **[Global] Automated traffic routing away from datacenter failure**

We can detect and route away from abrupt failures in **less than two minutes**. Our routing layer will stop routing clients to that datacenter and route them elsewhere.



Availability of service

“Availability is uptime. At any time I want to use a service, what is the confidence I can use it?” Dr Paddy Byers, CTO of Ably.

Ably is meticulously designed to be elastic and highly-available, providing the uptime and scale required for stringent and demanding realtime requirements. Our mathematically grounded design means we can transparently share operating boundaries we monitor to ensure capacity and therefore availability, helping you understand the type of scale and elasticity capable with Ably. We can also legitimately offer a 99.999% uptime SLA.

- **50% global capacity margin for instantaneous surge**

Ably operates at internet-scale, so our normal dimensions for capacity are already large. Regardless, we operate at 50% capacity margin so we can elastically deal with instant surges in demand and continue to be available in the event of AZ failure.

- **Connection capacity can double every 5 mins, halve every 10 mins & Channel capacity can double every 10 mins, halve every 20 mins**

Ably can react to changes and elastically scale beyond instant surge capacity. But we must maintain state in all new areas when scaling. To do this and allow the system to keep up as it scales, we constrain the ability of the system to double in capacity.

- **DoS: Layer 3, 4 and 7 defence in our edge network**

Ably has mechanisms to defend against DoS vectors across different layers. This includes Layer 7 ‘attacks’ that might be legitimate operations at unsustainable rates.

- **Max number of channels, throughput, and connections: limitless**

Ably can scale limitlessly. We achieve this with channel sharding, a mechanism to facilitate horizontal scaling. Each channel has limited capacity, but you can allocate as many channels as you need for your scale. For example, HubSpot employs over 500m channels per day. Unlimited connections includes fannout to millions of subscribers over a handful of channels, or one-to-one connections over individual channels.

- **99.9999% global service availability**

Ably is designed around the statistical probability that service availability will be 99.9999% (6x9s). To account for real-world behaviour, the lowest SLA we design around and commercially offer is 99.999% (5x9s).



Companies that depend on Ably

HubSpot

Bloomberg

yahoo!

OfferUp

Capgemini



NOKIA

Ably is an enterprise-ready pub/sub messaging platform delivering billions of messages everyday. Developers build on Ably to efficiently design, quickly ship, and seamlessly scale critical realtime functionality like HubSpot's live chat, in-play scores for millions of Australian Open fans, and realtime transit updates for three million Chicagoans without the engineering distraction, neverending project timelines, or runaway infrastructure costs.

We're the only pub/sub messaging platform delivering realtime functionality to more than 50 million daily end-users who are never more than 100ms away from our global network. We provide guarantees for message ordering and delivery to overcome limitations of unreliable network conditions. All without sacrificing security, fault tolerance, or uptime.

Learn more at ably.com or contact us ably.com/contact.

