Contents

- At some point in time, every organization realizes that its slowing down.
- If we've internalized essentialist bullshit about "B players hiring C players", maybe we get nervous about The Bar and hiring.
- If were anxious and uncertain, we might get religion about Agile or Scrum or whatever.
- These approaches rarely yield results.

1 Our Dogmatic Slumber

- Most explanations of organizational success or failure are crap. Emic accountsi.e., those from within the organizationare limited to those concepts and narratives which exist within the organization.
- If we take the etic perspectivei.e. from outside the organization-we can see that emic explanations for an organizations success or failure have readily available counterfactuals in other organizations. If agile, flat organizations, code reviews, monorepos, open offices,fancy type systems, etc. were actually the causal factors theyre purported to be, then why do so many organizations adopt those practices without success? Why are there other successful organizations which lack those practices? How can we tell the difference between *cum hoc, ergo propter hoc* just-so stories and actual causal factors?
- are there necessary, a priori truths of organizational performance?

2 Corporate America's Next Top Model

- If you squint hard enough, **an organization doing work is just an incredibly complex, dynamic, distributed, parallel process.**
- we can sketch out the *boundaries* of what an organization is capable of and the dynamics of that as it grows.
- What happens inside those boundaries is a matter of execution and effort; what happens outside those boundaries is impossible.

3 The Ceiling is Low

- **The work capacity of an organization scales, at most, linearly as new members are added.**
- Each new member in an organization adds a constant number of possible work hours to the total possible work hours of the company's existing employees. Amdahl's Law states that given a fixed task, a parallel solution utilizing N processors will run faster than a sequential solution by *at most* a factor of N.
- As parallel resources are added, the total time spent in the parallelizable portion of the task amortizes to zero; in contrast, the total time spent in the sequential portion of the task never drops below a floor value.
- Our intuition tells us that larger organizations do exhibit superlinear behaviors, but this literally cannot be the case if hiring is the only variable in the equation. Therefore, our only hope for superlinear productivity lies in changing the task which is being executed. Thankfully, work capacity is not the same as productivity.
- As an organization hires more employees, work on productivity improvements must be a constant priority.
- Internal tooling, training, and services must be developed and fielded to ensure that all members are able to work on problems of continuously increasing impact. **The ceaseless pursuit of force multipliers is the only possible route to superlinear productivity improvements as an organization grows.**

4 The Floor is Lava

- **Contention costs grow superlinearly as new members are added.**
- Parallel solutions to tasks are rarely perfectly concurrent (and indeed, such tasks are rightfully called "embarrassingly parallel"), and often require some sequential critical sections. The line of CPUs or people waiting to enter a critical section can be modeled as a queue, which allows us to use queuing theory to understand how the queue's cycle time changes as the queue size grows.
- If we model the line for a sequential section as a $G/G/1$ queue, which is to say, without making any assertions about the arrival process or service time distribution but assuming a single queue server (i.e. only one CPU or person can hold the lock), we arrive at Kingman's Formula for the mean wait time: $\mathbb{E}(W_q) \approx (\frac{\rho}{1-\rho})(\frac{{c_a}^2+{c_s}^2}{2})\tau$
- Notably, the wait time of a queue increases non-linearly with respect to ρ (utilization) and quadratically with respect to c_a (the coefficient of variation for arrivals) and c_s (the coefficient of variation for service times). (This is the quantified form of the intuition that queues are either empty or overflowing.)
- The non-linearity of this should give us pause, as increasing the number of people contending for a shared resource is the same thing as increasing ρ. If contention on those resources is unmanaged, organizational growth can result in catastrophic increases in wait time. At some point, adding new members can cause the organization's overall productivity to decrease instead of increase, as the increase in wait time due to contention is greater than the increase in work capacity. (This is the organizational version of the latency spikes we see as servers become overloaded.)
- A commonly applied but rarely successful strategy is using external resources–e.g. consultants, agencies, staff augmentation–as an end-run around contention on internal resources. While the consultants can indeed move quickly in a low-contention environment, integrating their work product back into the contended resources often has the effect of ballooning c_s (the variation of service times, or how long a critical section is held). This produces a quadratic spike in wait times which increases utilization which in turn produces a superlinear spike in wait times.
- Successful strategies for reducing contention include increasing the number of instances of a shared resource

  ...

  and developing stateless heuristics for coordinating access to shared resources.
- **Staffing highly sequential efforts as if they were entirely parallel leads to catastrophe.**

5 Hell is Other People

- **Coherence costs grow quadratically as new members are added.**
- A group of 3 has 3 dyads; a group of 4 has 6; a group of 5 has 10; a group of N people has $\frac{N^2-N}{2}$ possible dyads.
- If the relative percentage of people who need to talk to each other to get something done stays constant as the organization grows (i.e. $x\%$ of all dyads), **the total time spent communicating will grow quadratically as the work capacity of the organization grows linearly.**
- We can consider group meetings as a batching strategy to reduce the number of entities involved in point-to-point communications, but the effectiveness of this strategy depends heavily on the relative overlap of groups and the group structures.
- The only scalable strategy for containing coherence costs is to limit the number of people an individual needs to talk to in order to do their job to a constant factor.
- Each additional person or group in a responsibility assignment matrix geometrically increases the area of that matrix. Each additional responsibility assignment in that matrix geometrically increases the cost of organizational coherence.
- It's also worth noting that these pair-wise communications don't need to be formal, planned, or even well-known in order to have costs. Neither your employee handbook nor your calendar are accurate depictions of how work in the organization is done. Unless your organization is staffed with zombies, members of the organization will constantly be subverting standard operating procedure in order to get actual work done. Even ants improvise. An accurate accounting of these hidden costs can only be developed via an honest, blameless, and continuous end-to-end analysis of the work as it is happening.

6 Principles From Beyond Space And Time

6.1 Keep the work parallel, the groups small, and the resources local

- If the organization's intent is to increase value delivery by hiring more people, work efforts **must** be as independent as possible. Leaders should develop practices and processes to ensure that the work efforts which their strategies consider parallel are actually parallel.

6.2 Prioritize the development of force multipliers

- If an organization is largely working on the same types of problems it was in previous years, it's cause for concern.

6.3 If possible, factor work products into independent modules; if not, grow slowly and optimize

- If your work product–e.g. codebase, documents, etc.–can be factored into independent modules, do so. The key word there is *independent.* Slicing your shit up into a hundred microservices will not help you if everyone needs to change ten of them to get anything done.

6.4 Scale organizational efforts across a portfolio of synergistic products

- Most smart businesses start out with a single product. They go long on their product hypothesis, put their eggs in a single basket, and swing for the fences. If they're lucky enough to get traction, they double down on this.
- organization leaders should keep the development of a product portfolio as an explicit goal. Feature or product ideas which are complementary to the organization's overall business strategy but don't naturally coexist with the main product can be developed as separate products by independent teams.
- As a concrete example of the virtue of a product portfolio, imagine Amazon Web Services as a single product, staffed by a hundred thousand doomed souls. [...] Such a creature would implode under its own weight.
- Instead, Amazon Web Services is a portfolio of synergistic products.

6.5 Keep responsibility assignment matrices small, sparse, and local

- specialization is often critical for building internal economies of scale, but the formalization of new constituencies should be kept in check.
- Where a matrix indicates a high-touch relationship between two groups [...], efforts should be made to reduce the cost of that interaction by colocating their members (e.g. embed a lawyer with the engineers)

6.6 Prioritize asynchronous information distribution over synchronous

- A significant source of failure demand for meetings and status updates is the desire of organizational leaders to keep abreast of who's doing what.
- A better model for staying informed of developments as the organization scales is for groups to publish status updates as part of the regular cadence of their work. Leaders can asynchronously read these updates and, should the need arise, initiate additional, synchronous conversation to ask questions, provide feedback, etc.
- Synchronous meetings should be reserved for low-latency collaboration on complex issues; likewise, collaboration should be reserved for synchronous meetings.

6.7 What happens inside the boundaries is important

- Companies are groups of people being compensated for having to spend some of their finite lifetimes not being with their partners, children, pets, or super weird hobbies. They deserve to be members of organizations which honor that time by ensuring that their work has value and meaning.