Teaching Case

Viral Scalability - Coping with Sudden Demand Swings

Paul D. Witman
pwitman@callutheran.edu
School of Management
California Lutheran University
Thousand Oaks, CA, USA

Abstract

The rapid emergence of the novel coronavirus and its impact on human behavior provoked dramatic increases in human usage of a variety of systems. These increases had the potential to stress the scalability of the systems, testing whether the system owners had designed and built those systems to cope with sudden changes in demand. This case invites students to study a variety of different types of systems, and to analyze the means by which they can or did achieve scalability, and efficiently serve their customers.

Keywords: Scalability, coronavirus, cybersecurity, usability

1. OVERVIEW

The emergence of the novel coronavirus sparked a pandemic that rapidly upended many aspects of modern life. Over the course of just a few weeks, offices and schools of all types emptied out, their activities stopped or moved to various online formats. Retail stores and services closed, with many of those activities moved to e-commerce models. The shift was sudden and substantial, and tested the ability of systems and services to rapidly adjust to the increases in demand.

Organizations that were prepared, or made substantive changes quickly, for this sudden demand change will undoubtedly fare better in the weeks and months to come. Organizations that failed to scale quickly suffered reputational and other damage, and may face difficulty in recovering from those failures.

2. SCALABILITY DETAILS

Scalability is a "desirable attribute of a network, system, or process …, to process growing volumes of work gracefully, and/or to be susceptible to enlargement" (Bondi, 2000). Put another way – will the system or process continue to work well as demand for its services rises? Is "work well" a well-defined concept? How is it measured? What aspects of a system need to be scalable? How is scalability achieved?

We will use the real-world example of a grocery store to consider how these concepts relate, and then apply the concepts to technological systems.

Measuring scalability

Scalability is assessed by measuring various performance metrics for a system – throughput, storage capacity, client satisfaction, etc. As an example from the grocery store industry, we might choose to gather metrics such as:

- Shopper capacity – how many customers can safely be in the store at once?
- Throughput – how many people can find their needed groceries, check out, and depart in a given period of time?
- Speed of checkout – how long does it take the average customer with a particular collection of items to check out of the store?

- Sales volume – is there sufficient shelf and storeroom capacity to keep the shelves stocked?
- Resupply – is the entire supply chain able to keep the products in stock, in stores, distribution centers, warehouses, etc.?

Types of scalability

Bondi (2000) identifies four basic types of scalability:

- Load scalability – the ability to gracefully service increasing volumes of work without excess delay or resource usage. Example: number of people checking out of the store in a given period of time.
- Space scalability – the ability to service increasing volumes of work without running out of short-term or long-term storage. Example: Volume of product stored on shelves or in warehouse.
- Space-time scalability – the ability to continue to work in a time-efficient way even with a much larger volume of storage. Examples: Moving shelf re-stocking to night-time hours to avoid disrupting shoppers; making the store much larger to allow shoppers to spread out.
- Structural scalability – the ability of the system's design choices to support scalability requirements. Example: Are the shelves big enough to hold a day's demand for certain products before the next re-stocking? Are the aisles wide enough to permit social distancing?

In addition, some additional concepts apply more specifically to technological systems, but also to physical systems.

- Distance scalability – the system works well over short or long distances. Example: How far is the warehouse from the store? How quickly can a truck get to the store with an urgent request?
- Speed/distance scalability – the system works well over short or long distances, and at high or low speeds.

How to achieve scalability

Scalability is accomplished through design and implementation choices, most commonly through two general approaches: horizontal and vertical scalability. Horizontal scalability refers to the ability to have multiple of the same components able to do work in parallel – such as having multiple checkstands at the grocery store. Vertical scalability refers to switching to components that are individually higher in capacity – for example, using self-checkout to allow one staff member to supervise four customer checkouts at the same time.

Scalability also requires architectural and engineering choices to allow the various parts of a system to scale smoothly with one another. There's little value to the customer if we can check them out quickly, but the products they need are not available on the shelves.

Scalability in cloud services

Amazon Web Services (AWS) is a highly visible cloud service provider, and part of their value proposition is the ability to readily scale the capacity for a customer's workload. Morgan (2014) describes the architectural and implementation choices made by AWS to allow for rapid scalability. Those choices include:

- A range of geographic data center locations to address distance scalability.
- Advance planning and capital spending to keep server and network infrastructure ahead of customer requirements. An individual AWS data center is reputed to have 50,000 or more servers, which are partitioned with virtual machine capabilities to allow most efficient use.
- Operational spending to provide sufficient Internet service capacity to each data center.
- Capital and operational spending to provide the power and cooling infrastructure to service all of the equipment in each data center.

Amazon also makes extensive plans for scalability and reliability, including organizing its many data centers into Availability Zones. This enables customers to spread their workload not only among enough servers to carry the load, but also among data centers that are close enough to readily keep data synchronized between them.

3. SCALABILITY SCENARIOS

This section documents a range of different scenarios, across a variety of industries, some technology-centric, others less so. Each provides an opportunity to consider the various dimensions of scalability, and to analyze what had to have happened to make that scalability work. Each also provides an opportunity to consider potential problems that still exist, and what problems may still be occurring, but at low frequencies.

Online Conferencing

One very apparent effect of the pandemic was the sudden transition from in-person activities to online video conferencing. One major player in this space, Zoom, reported that their usage grew from 10 million active users to 300 million active users over the months of March and April, 2020 (Grant, 2020). Downdetector.com reports relatively stable numbers of problem reports over those same months, usually fewer than 100 reports per day. One analysis of Zoom's prior history (pre-pandemic) showed that Zoom had not intended to be operated at this massive scale (Bennett and Grant, 2020).

That said, Zoom's growth was not without problems, though perhaps not classically scaling problems (Keck, 2020; Paul, 2020; Zakrzewski and Riley, 2020b). Zoombombing became part of the vocabulary in the US and elsewhere, used to describe rogue users. These (sometimes uninvited) users behaved badly in meetings, sharing unwanted screen views and making noise. In response, Zoom issued updated software that strengthened default security choices.

And of course, this same phenomenon was seen in other countries. For example, after Chinese workers returned to work after an extended Chinese New Year holiday, user counts at conference site DingTalk grew from 26 million to 150 million from January 1 to February 21. WeChat Work's user count more than doubled, from 5.6 million to 13 million, in the same period. DingTalk also reported that there were over 200 million users (some of them the same person in multiple meetings) connecting to meetings on that first day of work, February 3, 2020 (Zhijie and Xin, 2020).

Additional details about online conferencing scalability and its dependencies may be found in the References, such as Baker (2020) and Bennett (2020).

Questions:

- What are the various elements that had to work well for Zoom (or other such services) to scale as readily as it did?
- Why do you think there was a sudden uptick in apparent Zoom security issues, when the software had these issues before the usage spike?
- Scalability can apply to sudden decreases in utilization as well. What can Zoom or other such services do to prepare for sudden decreases such that their cost model scales up and down to maintain profitability?
- Based on this example, what general observations about scalability can you make?

Internet Capacity

The sudden increase in the number of people staying home resulted in increases in, among other things, video conferencing as well as entertainment video streaming. Data consumption in some parts of Europe grew by 30% over a short period of time, with streaming video (pre-COVID) estimated to be 60% of total consumer network traffic. A sudden increase in streaming video, coupled with a huge rise in video conferencing, threatened to cause Internet outages (Baker, 2020).

As a result, Netflix and YouTube, two of the largest streaming video providers, agreed to reduce the quality of their video deliveries in order to reduce data volumes (Gold, 2020). This was expected to reduce total data usage by these two providers by 25%. Other providers, such as Amazon Prime Video and Facebook's video streaming services made similar changes to reduce data usage.

Questions:

- Why does reducing video quality impact data volumes?
- Does reducing the delivered video quality necessarily impact the user's perceived video quality? Why or why not?
- What are some differences between live-streamed video or video conferencing, relative to recorded video? Which one is more likely to be negatively impacted by Internet capacity issues?
- Based on this example, what general observations about scalability can you make?

Online Fitness Services

Fitness services (gyms, spas, exercise studios, and the like) were all quickly impacted by COVID-related shutdowns (Newcomer, 2020). Peloton, an exercise equipment and services company, reported hosting its largest-ever online class, with 23,000 active participants.

The authors spoke with one large service provider (WellnessCo) to that industry, who reported that of the 60,000+ customer locations (usually run by a small business owner) worldwide, about 50,000 of them closed within a several week

period. The company was able to see the pattern developing first in Asia and Europe, and then to anticipate the changes coming to the US and Latin American markets.

In a bid to quickly restructure their business for a long period without in-person customers, the company built a series of virtual wellness applications leveraging Amazon cloud services. They include online exercise classes for small groups using Amazon Chime, and large-group game and competition tools based on Amazon Twitch. They built and launched the services over just a few weeks and had 240,000 new classes in operation each day within four weeks of product launch.

Questions:

- What might be the motivation for WellnessCo to use Amazon's cloud offerings to build its new products?
- Certainly, as demand rapidly increased as lockdowns spread, scaling up is important. Do you anticipate a need for the company to be able to scale down? Why?
- How would you anticipate consumer behavior to change over time as the pandemic ends, a vaccine takes hold, etc.? Will all consumers revert to the in-person fitness classes? Why?
- Based on this example, what general observations about scalability can you make?

Mail-in Voting

The contagion risk created by the pandemic has prompted many US states to change their voting procedures. In many cases, states have proposed moving to an all-mail ballot procedure. In this model, voters receive a ballot package in the mail, and are requested to vote their preferences and mail the completed ballot back to election officials. The vote may involve punching out a hole in the ballot, marking a spot on the ballot with a pen, or other means to convey the voter's decision.

Mail-in ballots are already used in many states for voters who for some reason cannot vote in person in their local area, so the basic infrastructure already exists. However, moving all voting to a mail-in model requires a number of things to happen in much higher volumes, from packaging and mailing ballots, to verifying voter identities, to counting the ballots once returned (often using different machines for the in-person ballot counting (Marks and Riley, 2020).

Questions:

- What are the various aspects of the voting system that will need to scale differently to support an entirely or mostly vote-by-mail election? Include systems controlled by elections officials, and those outside of their control. These could be human, mechanical, or electronic systems.
- Where are the potential bottlenecks in the end to end system? What could be done to mitigate the risk of those bottlenecks?
- Based on this example, what general observations about scalability can you make?

Economic Stimulus Checks

In March of 2020, the US government passed legislation that would deliver money to most individual taxpayers, as a support for individuals and the economy. The funds were slated to be delivered electronically to the account used for the individual's tax payment in the most recent tax year. The US government provided some guidance as to the timing, but it was not very specific, leaving users to wonder when their particular payment would arrive.

As the deadline for these deposits approached, large numbers of online banking users logged in, causing unexpected surges of activity for many online banking web sites and apps. This caused slow response times for some, and caused bank web sites and apps to fail entirely for others (Mak, 2020; Rayome, 2020).

Questions:

- Thinking about your own usage of online banking, what types of routine events would cause a surge in online banking activities?
- Given that banks generally seem to plan and execute well enough to address those routine events, why do you think the stimulus check logins created such problems?
- Research the banking system, and investigate the technological components involved in supporting online banking. Draw a diagram of the components and identify potential bottlenecks.
- Based on this example, what general observations about scalability can you make?

Streaming Church Services

Social media and online services of various forms are often used by non-profit organizations to accomplish their organizational missions (Witman, 2013). In a COVID-19 world, religious organizations were largely forced to move from traditional in-person meetings and worship services to pre-recorded and live-streamed worship delivery.

One church in Southern California chose Facebook Live as its delivery media. Their first week of services went very smoothly, with no significant technological hiccups. See Figure 1, Appendix, for its beginning technology design.

The second week was a bit more challenging. They made seemingly minor technical changes, adding an additional laptop connected to WiFi AP #1, livestreaming content from Facebook, and responding to Chat messages, as well as adding an additional camera for picture in picture. The result was significant pauses – 30-90 seconds – where both video and audio froze for the remote viewers. See Figure 2 for details.

In the third week, they moved the added laptop from the second week to a separate WiFi Access Point to reduce stress on AP#1. Watching a livestream on Facebook reportedly consumes 1-8Mbps; Facebook specifications call for a minimum 5Mbps uplink speed. See Figure 3 for details.

Third week results were more stable, and no new systemic issues have been reported since. It is worth noting, though, that this type of problem can be difficult to diagnose. There might be bottlenecks in the church's own infrastructure (routers, etc.), and in the local internet service provider. From the view of the end user, glitches can be triggered by the video streaming service, the user's viewing device, or any network component in between. Not a simple problem to trace!

Questions:

- Based on the information provided, what seems to be the likely causes for the issues discovered in Week 2?
- What steps, if any, could the church's tech team have done to have caught the Week 2 issues before the event took place publicly?
- How could the church ensure, before the time of a live-stream, that the technology was all working correctly and scalably?
- Based on this example, what general observations about scalability can you make?

Unemployment Insurance Claims

As businesses closed due to the pandemic, many employees were furloughed. Unemployment insurance is a mechanism used in the US to provide some level of income replacement for workers who are temporarily out of work. In many cases, the process of applying for that assistance is intended to be done online, and sometimes only online. The process is administered by the state in which the person lives.

In addition, some states required benefit recipients to log in each week to certify that they remained unemployed and thus eligible for benefits. Many states experienced problems with their online systems, with slow responses, failed application submissions, and other errors preventing access. (Fineout and Caputo, 2020; Zakrzewski and Riley, 2020a).

Questions:

- Given the nature of unemployment insurance as a government-operated social service (rather than a for-profit company), what is the economic or other incentive that would motivate the service to plan appropriately for scalability?
- Some states require all recipients to log in each week. Why does this process affect unemployment website scalability?
- Based on this example, what general observations about scalability can you make?

Family/Friend Connections from Isolation

One unexpected benefit that comes with easy-to-use technology is that people who had not previously used videoconferencing were able to connect in ways that were previously perceived to be out of reach, either technically or in terms of costs. One student at a southern California university reported that his family was able to connect with an aging grandmother, including many family members from across the US and Latin America. They held daily prayer meeting with the grandmother, live online with many family members, a level of engagement not imagined in a pre-COVID world.

In a more structured fashion, senior centers and arts and other organizations are providing online

classes, exercise groups, and social gatherings for older individuals (Finn, 2020).

Questions:

- What happened along with the sudden increase in Zoom and other video conference utilization to enable this sort of additional services to be available to older individuals? Think about things like the broad visibility of conferencing, its ease of use, etc.
- Why do you think online services like Zoom did not get more of this type of usage before the COVID crisis?
- Based on this example, what general observations about scalability can you make?

4. FOLLOW-UP QUESTIONS

These are some follow-up questions to provoke further research and exploration. Your instructor may have other questions, and we encourage you to develop your own questions as well.

- What other examples have you seen where the ability to scale a system has been important? What happened? Why was it important?
- Thinking about the range of different systems and processes noted here, and those that scaled better than others, are there any common aspects that you can identify about successes vs. challenges in systems and organizations?

5. FOLLOW-UP RESEARCH

COVID-19 Vaccine Supply Chain

At the time of this writing, a coronavirus vaccine is yet to be available. Even when the scientific work of designing and testing the vaccine is complete, an enormous logistical challenge lies ahead to get that vaccine (or those vaccines) to as many people and parts of the world as possible, as quickly as possible (NBC News, 2020; Owermohle, 2020). Even things as mundane as the containers for vaccine doses have to be manufactured (billions of them), and vaccines usually have to be shipped under tight constraints of timing and temperature control. Some vaccines have to be administered in two separate doses spaced appropriately apart as well.

Questions:

- What are the challenges in delivering and administering a vaccine like this?
- What are the risks if the supply chain is not managed well? How might they be mitigated?
- Are the risks and challenges likely to be the same all around the world? Why might we need to do location-specific planning for vaccine delivery?

Other examples …

Doubtless you have had your own opportunity to observe scalability in action, or in failure, either due to the COVID-19 virus, or due to other factors.

- Document the system or process that needed to scale, perhaps with a diagram.
- What were the critical points that had to scale well?
- Where did problems crop up?
- How were the problems resolved?

6. CONCLUSIONS

Scalability is an important factor to consider in building any system, and there are often many components that need to interact efficiently to achieve scalability. This applies not just to technological systems but to all organizational functions, to ensure that the system can handle sudden increases in demand, as well as cost-effectively handle sudden or steady decreases in demand. It is instructive for business and information technology students to study the scalability of systems in order to prepare their organizations, and their technology, for these inevitable (though sometimes unprecedented) changes.

7. ACKNOWLEDGEMENTS

The author appreciates the support and insights of various research subjects who provided interviews for this paper. The author also appreciates the productive feedback provided by Bill Naylor, and by the reviewers and conference chairs.

8. REFERENCES

Baker, M. A. (2020). Working Together to Keep America Connected. Retrieved from https://www.ctia.org/news/blog-working-together-to-keep-america-connected

Bennett, D., & Grant, N. (2020, April 9). Zoom Goes from Conferencing App to the Pandemic's Social Network. *Bloomberg BusinessWeek*.

Bondi, A. B. (2000). *Characteristics of scalability and their impact on performance.* Paper presented at the Proceedings of the 2nd international workshop on Software and performance.

Fineout, G., & Caputo, M. (2020, April 3). Florida's nightmare with unemployment could hurt Trump. Retrieved from https://www.politico.com/states/florida/story/2020/04/03/its-a-sh-sandwich-republicans-rage-as-florida-becomes-a-nightmare-for-trump-1271172

Finn, C. (2020, April 6). How tech is helping elderly fight coronavirus lockdown loneliness. Retrieved from https://www.aljazeera.com/news/2020/04/tech-helping-elderly-fight-coronavirus-lockdown-loneliness-200402185238544.html

Gold, H. (2020, March 20). Netflix and YouTube are slowing down in Europe to keep the internet from breaking. Retrieved from https://www.cnn.com/2020/03/19/tech/netflix-internet-overload-eu/index.html

Grant, N. (2020, April 22). Zoom Daily Users Surge to 300 Million in Coronavirus Lockdown. Retrieved from https://www.bloomberg.com/news/articles/2020-04-22/zoom-daily-users-surge-to-300-million-despite-privacy-woes

Keck, C. (2020, April 22). Zoom Rolls Out Security Updates Following Zoombombing and Glaring Security Failures. Retrieved from https://gizmodo.com/zoom-rolls-out-security-updates-following-zoombombing-a-1843008545

Mak, A. (2020, April 15). Bank Websites Are Crashing Because Everyone Wants Their Stimulus Money. Retrieved from https://slate.com/technology/2020/04/coronavirus-stimulus-payments-checks-bank-irs-down.html

Marks, J., & Riley, T. (2020, March 27). Nationwide voting by mail will be a massive undertaking say those who've done it. Retrieved from https://www.washingtonpost.com/news/powerpost/paloma/the-cybersecurity-202/2020/03/27/the-cybersecurity-202-nationwide-voting-by-mail-will-be-a-massive-undertaking-say-those-who-ve-done-it/5e7ced39602ff10d49ad5a3b/

Morgan, T. P. (2014, November 14). A Rare Peek Into The Massive Scale of AWS. Retrieved from https://www.enterpriseai.news/2014/11/14/rare-peek-massive-scale-aws/

NBC News (Producer). (2020, June 12). What it will take to distribute a coronavirus vaccine to the masses. [Video] Retrieved from https://www.msn.com/en-us/video/tunedin/what-it-will-take-to-distribute-a-coronavirus-vaccine-to-the-masses/vi-BB15rCdP

Newcomer, E. (2020, April 24). Peloton Attracts a Record 23,000 People to Single Workout Class. Retrieved from https://www.bloomberg.com/news/articles/2020-04-24/peloton-attracts-a-record-23-000-people-to-single-workout-class

Owermohle, S. (2020, May 11). The 'biggest challenge' won't come until after a coronavirus vaccine is found. *Politico*. Retrieved from https://www.politico.com/news/2020/05/11/coronavirus-vaccine-supply-shortages-245450

Paul, K. (2020, April 2). 'Zoom is malware': why experts worry about the video conferencing platform. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2020/apr/02/zoom-technology-security-coronavirus-video-conferencing

Rayome, A. D. (2020, April 15). Online banking struggles as people check for coronavirus stimulus checks. *C|Net Personal Finance*. Retrieved from https://www.cnet.com/personal-finance/online-banking-struggling-as-people-check-for-coronavirus-stimulus-checks/

Witman, P. (2013). Social media for social value. *Computer, 46*(7), 82-85.

Zakrzewski, C., & Riley, T. (2020a, April 2). State unemployment websites are crashing amid record number of claims. *Technology 202*. Retrieved from https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/04/02/the-technology-202-state-unemployment-websites-are-crashing-amid-record-number-of-claims/5e84ee3e88e0fa101a758301/

Zakrzewski, C., & Riley, T. (2020b, April 3). Zoom chief Eric Yuan says he was not prepared for flood of security and privacy complaints. *Technology 202*. Retrieved from https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/04/03/the-technology-202-zoom-chief-eric-yuan-was-not-prepared-for-flood-of-security-and-privacy-complaints/5e860e3a602ff10d49adb7c8/

Zhijie, Y., & Xin, M. (2020, July 1). Homeward Bound. *ChinaNews, 143,* 41-43.

AUTHOR BIOGRAPHY

Paul D. Witman is a Professor in Information Technology Management at California Lutheran University, and Director of the School's graduate programs in Information Technology. His research interests include teaching cases, information security and privacy, the use of information related to vulnerable population groups, and the uses of social media in nonprofits and religious organizations.

.

ISSN: 2473-4901
v6 n5345

Appendices and Annexures

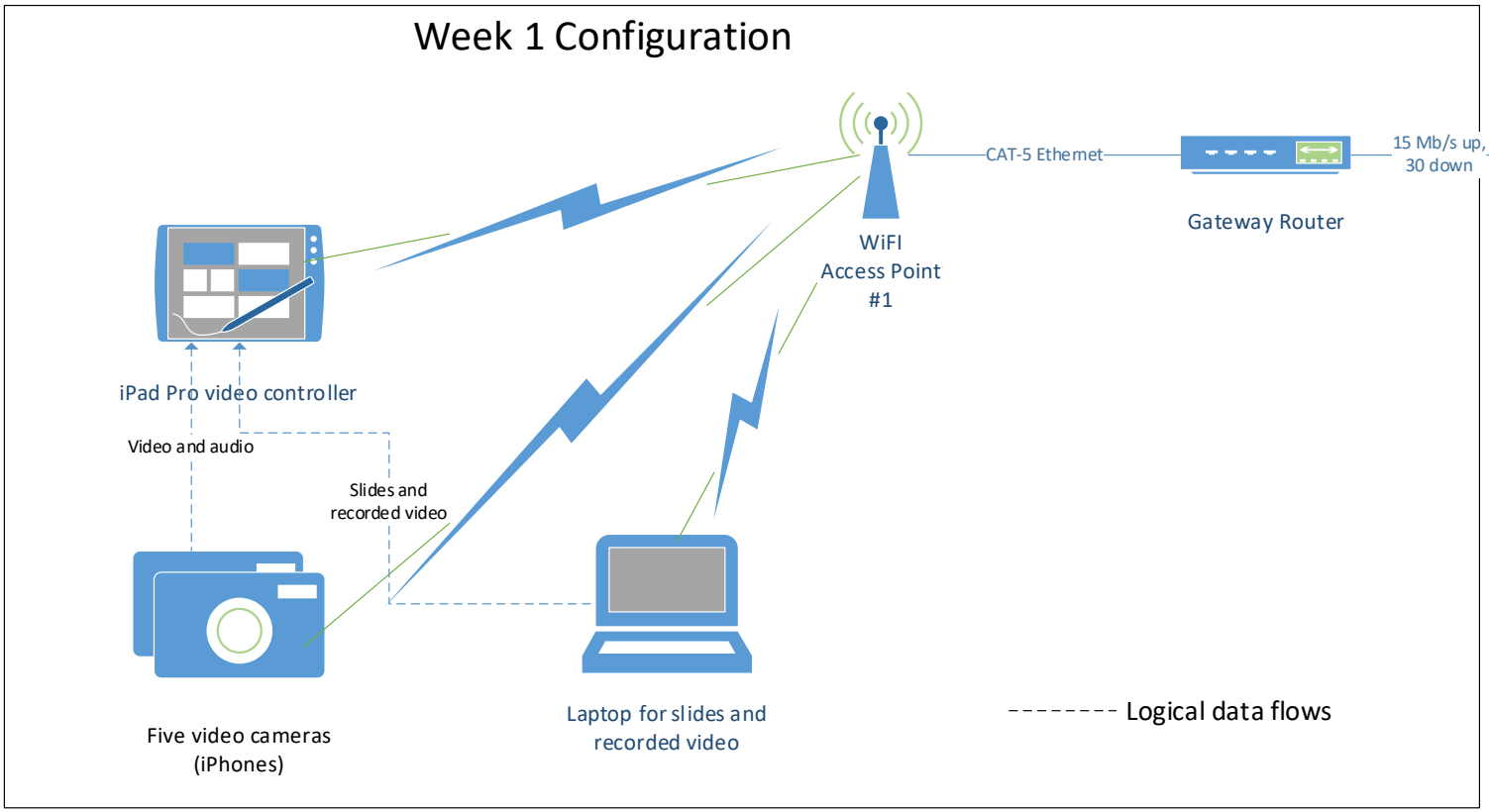

Figure 1 – Church's live-streaming setup, Week 1

ISSN: 2473-4901
v6 n5345

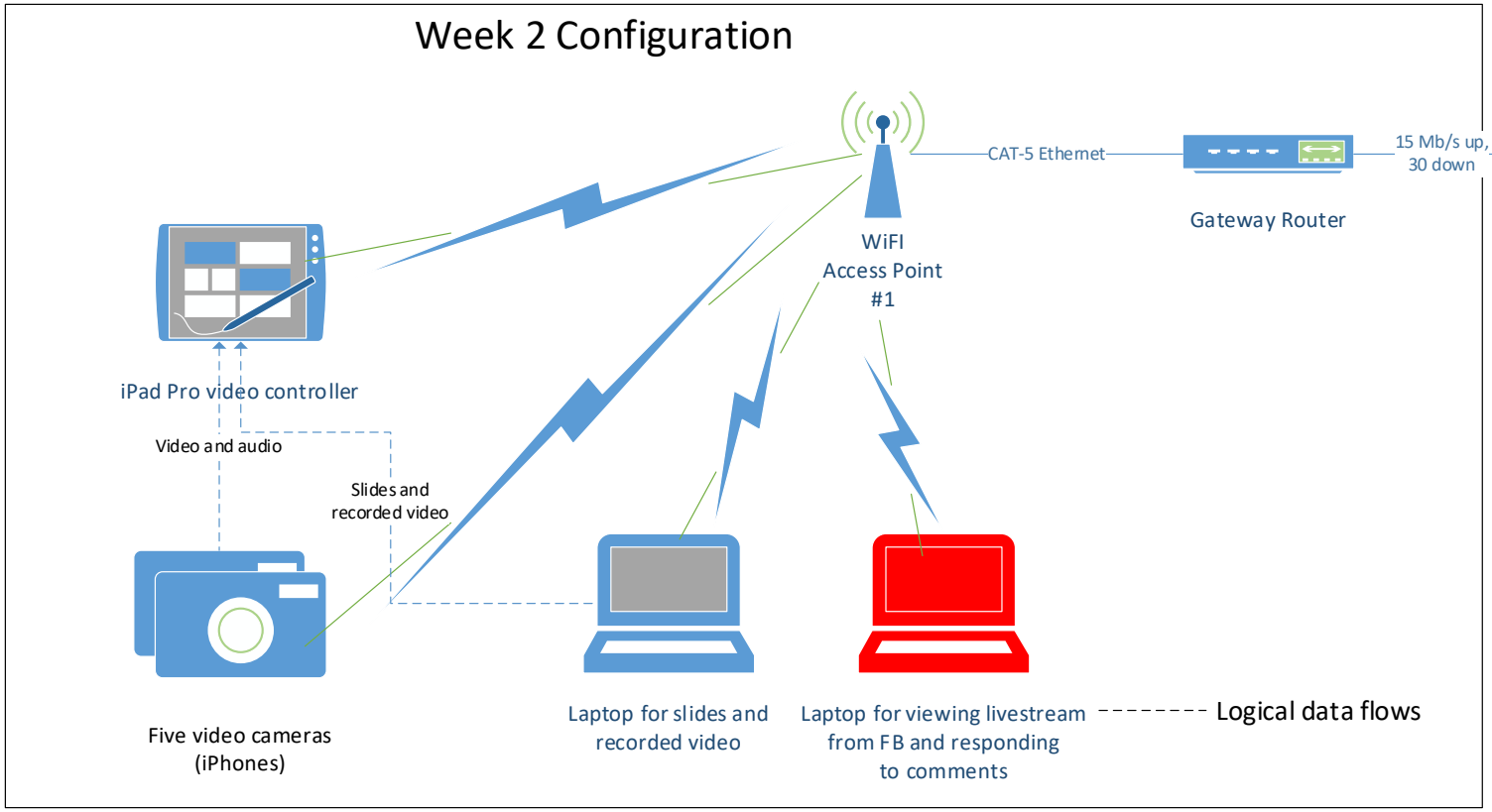

Figure 2 – Church's live-streaming setup, Week 2

ISSN: 2473-4901
v6 n5345

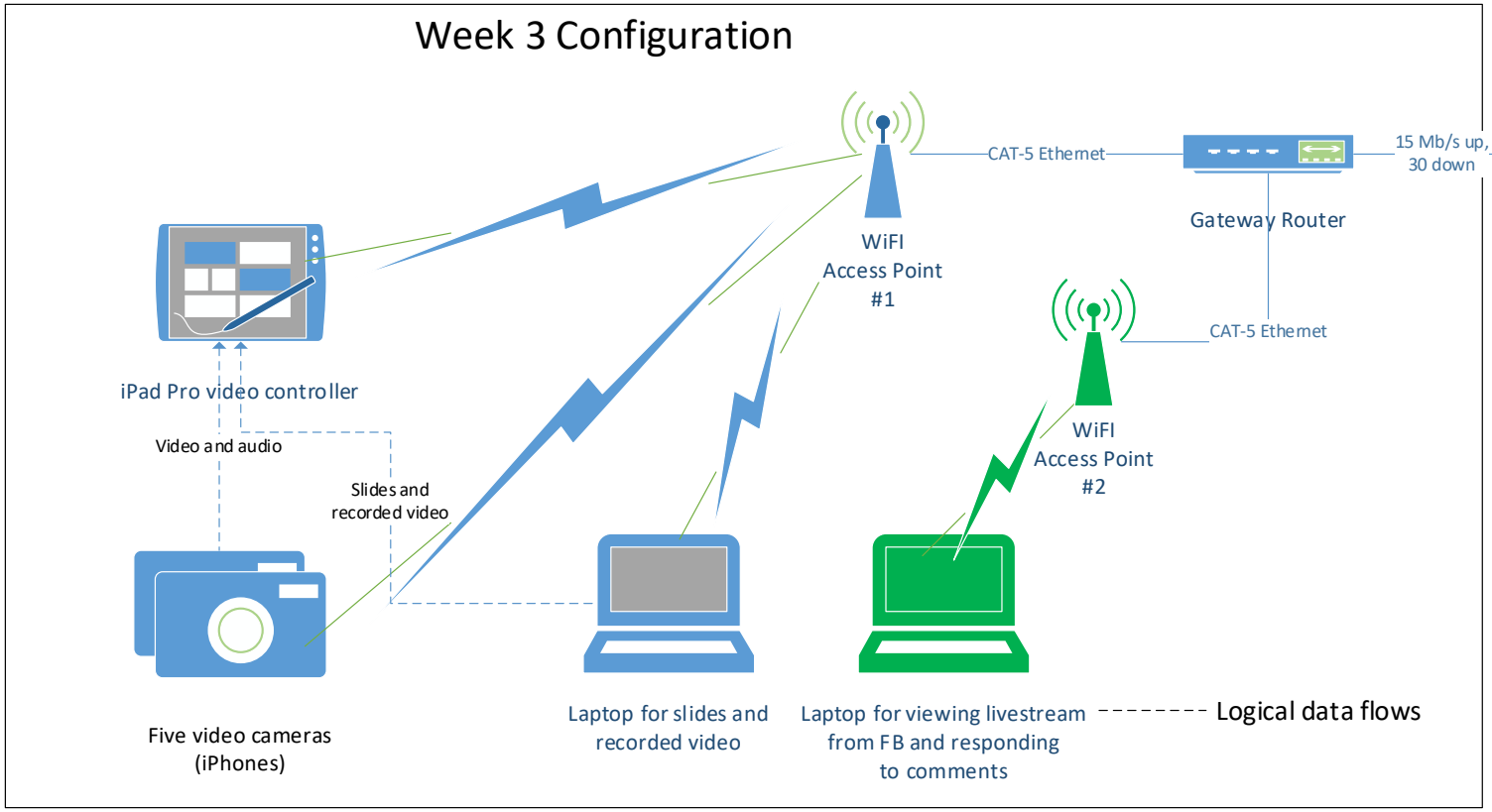

Figure 3 – Church's live-streaming setup, Week 3