BeyondCorp

A New Approach to Enterprise Security

RORY WARD AND BETSY BEYER

Rory Ward is a site reliability engineering manager in Google Ireland. He previously worked in Ireland at Valista, in Silicon Valley at AOL, Netscape, Kiva, and General Magic, and in Los Angeles at Retix. He has a BSc in computer applications from Dublin City University. roryward@google.com

Betsy Beyer is a technical writer specializing in virtualization software for Google SRE in NYC. She has previously provided documentation for Google Data Center and Hardware Operations teams. Before moving to New York, Betsy was a lecturer in technical writing at Stanford University. She holds degrees from Stanford and Tulane. bbeyer@google.com

Virtually every company today uses firewalls to enforce perimeter security. However, this security model is problematic because, when that perimeter is breached, an attacker has relatively easy access to a company's privileged intranet. As companies adopt mobile and cloud technologies, the perimeter is becoming increasingly difficult to enforce. Google is taking a different approach to network security. We are removing the requirement for a privileged intranet and moving our corporate applications to the Internet.

Since the early days of IT infrastructure, enterprises have used perimeter security to protect and gate access to internal resources. The perimeter security model is often compared to a medieval castle: a fortress with thick walls, surrounded by a moat, with a heavily guarded single point of entry and exit. Anything located outside the wall is considered dangerous, while anything located inside the wall is trusted. Anyone who makes it past the drawbridge has ready access to the resources of the castle.

The perimeter security model works well enough when all employees work exclusively in buildings owned by an enterprise. However, with the advent of a mobile workforce, the surge in the variety of devices used by this workforce, and the growing use of cloud-based services, additional attack vectors have emerged that are stretching the traditional paradigm to the point of redundancy. Key assumptions of this model no longer hold: The perimeter is no longer just the physical location of the enterprise, and what lies inside the perimeter is no longer a blessed and safe place to host personal computing devices and enterprise applications.

While most enterprises assume that the internal network is a safe environment in which to expose corporate applications, Google's experience has proven that this faith is misplaced. Rather, one should assume that an internal network is as fraught with danger as the public Internet and build enterprise applications based upon this assumption.

Google's BeyondCorp initiative is moving to a new model that dispenses with a privileged corporate network. Instead, access depends solely on device and user credentials, regardless of a user's network location—be it an enterprise location, a home network, or a hotel or coffee shop. All access to enterprise resources is fully authenticated, fully authorized, and fully encrypted based upon device state and user credentials. We can enforce fine-grained access to different parts of enterprise resources. As a result, all Google employees can work successfully from any network, and without the need for a traditional VPN connection into the privileged network. The user experience between local and remote access to enterprise resources is effectively identical, apart from potential differences in latency.

The Major Components of BeyondCorp

BeyondCorp consists of many cooperating components to ensure that only appropriately authenticated devices and users are authorized to access the requisite enterprise applications. Each component is described below (see Figure 1).

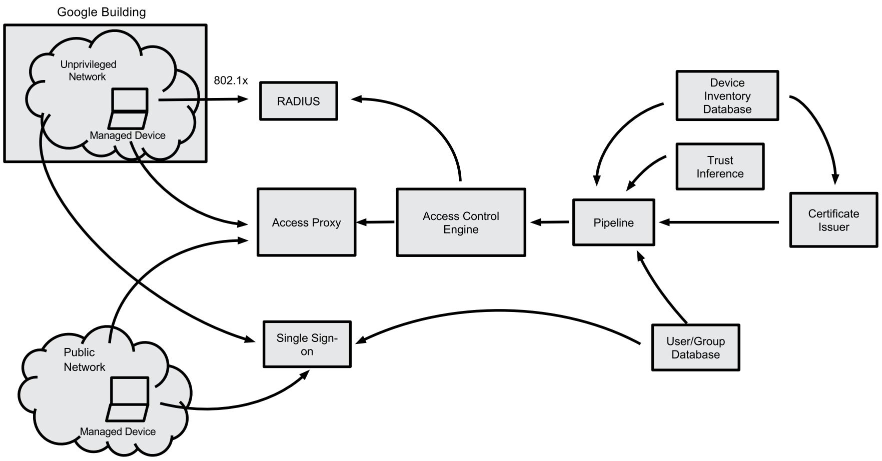

Figure 1: BeyondCorp components and access flow

Securely Identifying the Device

Device Inventory Database

BeyondCorp uses the concept of a "managed device," which is a device that is procured and actively managed by the enterprise. Only managed devices can access corporate applications. A device tracking and procurement process revolving around a device inventory database is one cornerstone of this model. As a device progresses through its life cycle, Google keeps track of changes made to the device. This information is monitored, analyzed, and made available to other parts of BeyondCorp. Because Google has multiple inventory databases, a meta-inventory database is used to amalgamate and normalize device information from these multiple sources, and to make the information available to downstream components of BeyondCorp. With this meta-inventory in place, we have knowledge of all devices that need to access our enterprise.

Device Identity

All managed devices need to be uniquely identified in a way that references the record in the Device Inventory Database. One way to accomplish this unique identification is to use a device certificate that is specific to each device. To receive a certificate, a device must be both present and correct in the Device Inventory Database. The certificate is stored on a hardware or software Trusted Platform Module (TPM) or a qualified certificate store. A device qualification process validates the effectiveness of the certificate store, and only a device deemed sufficiently secure can be classed as a managed device. These checks are also enforced as certificates are renewed periodically. Once installed, the certificate is used in all communications to enterprise services. While the certificate uniquely identifies the device, it does not single-handedly grant access privileges. Instead, it is used as a key to a set of information regarding the device.

Securely Identifying the User

User and Group Database

BeyondCorp also tracks and manages all users in a User Database and a Group Database. This database system tightly integrates with Google's HR processes that manage job categorization, usernames, and group memberships for all users. As employees join the company, change roles or responsibilities, or leave the company, these databases are updated. This system informs BeyondCorp of all appropriate information about users that need to access our enterprise.

Single Sign-On System

An externalized, single sign-on (SSO) system is a centralized user authentication portal that validates primary and second-factor credentials for users requesting access to our enterprise resources. After validating against the User Database and Group Database, the SSO system generates short-lived tokens that can be used as part of the authorization process for specific resources.

Removing Trust from the Network

Deployment of an Unprivileged Network

To equate local and remote access, BeyondCorp defines and deploys an unprivileged network that very closely resembles an external network, although within a private address space. The unprivileged network only connects to the Internet, limited infrastructure services (e.g., DNS, DHCP, and NTP), and configuration management systems such as Puppet. All client devices are assigned to this network while physically located in a Google building. There is a strictly managed ACL (Access Control List) between this network and other parts of Google's network.

802.1x Authentication on Wired and Wireless Network Access

For both wired and wireless access, Google uses RADIUS servers to assign devices to an appropriate network, based on 802.1x authentication. We use dynamic, rather than static, VLAN assignment. This approach means that rather than relying on the switch/port static configuration, we use the RADIUS servers to inform the switch of the appropriate VLAN assignment for the authenticated device. Managed devices provide their certificate as part of this 802.1x handshake and are assigned to the unprivileged network, while unrecognized and unmanaged devices on the corporate network are assigned to a remediation or guest network.

Externalizing Applications and Workflows

Internet-Facing Access Proxy

All enterprise applications at Google are exposed to external and internal clients via an Internet-facing access proxy that enforces encryption between the client and the application. The access proxy is configured for each application and provides common features such as global reachability, load balancing, access control checks, application health checks, and denial-of-service protection. This proxy delegates requests as appropriate to the back-end application after the access control checks (described below) complete.

Public DNS Entries

All of Google's enterprise applications are exposed externally and are registered in public DNS with a CNAME pointing the applications at the Internet-facing access proxy.

Implementing Inventory-Based Access Control

Trust Inference for Devices and Users

The level of access given to a single user and/or a single device can change over time. By interrogating multiple data sources, we are able to dynamically infer the level of trust to assign to a device or user. This level of trust can then be used by the Access Control Engine (described below) as part of its decision process. For example, a device that has not been updated with a recent OS patch level might be relegated to a reduced level of trust. A particular class of device, such as a specific model of phone or tablet, might be assigned a particular trust level. A user accessing applications from a new location might be assigned a different trust level. We use both static rules and heuristics to ascertain these levels of trust.

Access Control Engine

An Access Control Engine within the access proxy provides service-level authorization to enterprise applications on a per-request basis. The authorization decision makes assertions about the user, the groups to which the user belongs, the device certificate, and artifacts of the device from the Device Inventory Database. If necessary, the Access Control Engine can also enforce location-based access control. The inferred level of trust in the user and the device is also included in the authorization decision. For example, access to Google's bug tracking system can be restricted to full-time engineers using an engineering device. Access to a finance application can be restricted to full-time and part-time employees in the finance operations group using managed non-engineering devices. The Access Control Engine can also restrict parts of an application in different ways. For example, viewing an entry in our bug tracking system might require less strict access control than updating or searching the same bug tracking system.

Pipeline into the Access Control Engine

The Access Control Engine is constantly fed by a running pipeline that dynamically extracts information useful for access decisions. Among other factors, this information includes certificate whitelists, trust levels of devices and users, and inventory details about the device and the user.

An End-to-End Example

The Application

For this example, let us assume an application is to be taken BeyondCorp. The application is used by engineers to review source code, comment on the code, update the code, and, when approved by reviewers, submit the code. The application, codereview.corp.google.com, is restricted to full-time and part-time engineers from any managed device.

Configuring the Internet-Facing Access Proxy

The owner of codereview.corp.google.com configures the access proxy for the service. The configuration specifies the location of the back ends and the maximum traffic accepted by each back end. The codereview.corp.google.com domain name is registered in public DNS with a CNAME pointing to the access proxy. For example:

```
$ dig @8.8.8.8 codereview.corp.google.com

; <<>> DiG 9.8.1-P1 <<>> @8.8.8.8 codereview.corp.google.com
; (1 server found)
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 12976
;; flags: qr rd ra; QUERY: 1, ANSWER: 2, AUTHORITY: 0,
ADDITIONAL: 0

;; QUESTION SECTION:
;codereview.corp.google.com. IN   A

;; ANSWER SECTION:
codereview.corp.google.com. 21599  IN   CNAME
accessproxy.l.google.com.
accessproxy.l.google.com.    299    IN   A      74.125.136.129
```

```
;; Query time: 10 msec
;; SERVER: 8.8.8.8#53(8.8.8.8)
;; WHEN: Wed Aug 20 19:30:06 2014
;; MSG SIZE rcvd: 86
```

Configuring the Access Control Engine

The Access Control Engine provides a default rule that restricts access to full-time employees using a managed device. The owner of codereview.corp.google.com provides a more specific rule that further restricts access in two ways: to managed devices with the highest trust level, and to full-time and part-time engineers with the highest trust level.

An Engineer Accesses a Network

If the Network Is Located Outside a Physical Building Operated by the Enterprise: From a laptop provided by Google, an engineer accesses any WiFi network. For example, this network might be an airport WiFi network with a captive portal or a coffee shop's WiFi. There is no requirement to set up a VPN connection to the enterprise network.

If the Network Is Located in a Physical Building Operated by the Enterprise: From a laptop or desktop provided by Google, an engineer accesses the enterprise network. The laptop provides its device certificate in the 802.1x handshake with the RADIUS servers. As a valid certificate is provided, the laptop is assigned an address on the unprivileged network. If the device is not a corporate-issued laptop, or its certificate has expired, the device is assigned an address on a remediation network, which has very limited access rights.

Accessing the Application, Regardless of Network

From a corporate-issued laptop on a network, an engineer accesses codereview.corp.google.com. You can refer back to Figure 1 as a reference for the flow for this process.

1. The request is directed to the access proxy. The laptop provides its device certificate.
2. The access proxy does not recognize the user and redirects to the SSO system.
3. The engineer provides his or her primary and second-factor authentication credentials, is authenticated by the SSO system, is issued a token, and is redirected back to the access proxy.
4. The access proxy now has the device certificate, which identifies the device, and the SSO token, which identifies the user.
5. The Access Control Engine performs the specific authorization check configured for codereview.corp.google.com. This authorization check is made on every request:
   a. The user is confirmed to be in the engineering group.
   b. The user is confirmed to possess a sufficient trust level.
   c. The device is confirmed to be a managed device in good standing.
   d. The device is confirmed to possess a sufficient trust level.
   e. If all these checks pass, the request is passed to an appropriate back end to be serviced.
   f. If any of the above checks fails, the request is denied.

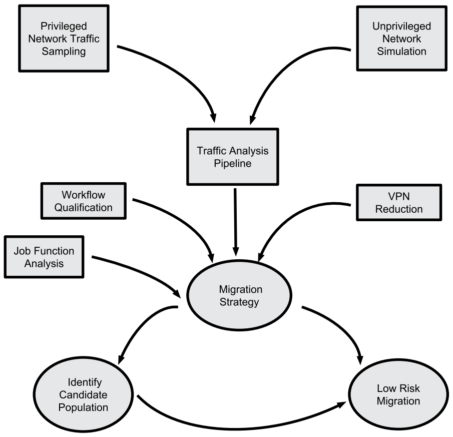
Privileged Network Traffic Sampling
Unprivileged Network Simulation
Traffic Analysis Pipeline
Workflow Qualification
VPN Reduction
Job Function Analysis
Migration Strategy
Identify Candidate Population
Low Risk Migration

Figure 2: Migrating to BeyondCorp

With this approach, we have rich, service-level authentication and authorization checks that are exercised on a per-request basis.

Migrating to BeyondCorp

Like virtually every other enterprise in the world, Google maintained a privileged network for its clients and applications for many years. This paradigm gave rise to significant infrastructure that is critical to the day-to-day workings of the company. While all components of the company will migrate to BeyondCorp, moving every network user and every application to the BeyondCorp environment in one fell swoop would be incredibly risky to business continuity. For that reason, Google has invested heavily in a phased migration that has successfully moved large groups of network users to BeyondCorp with zero effect on their productivity. The following section, represented by Figure 2, details some of the work we have done.

Workflow Qualification

All the applications used at Google are required to work through the access proxy. The BeyondCorp initiative examined and qualified all applications, which accomplish tasks ranging from the simple (e.g., supporting HTTPS traffic) to the more difficult (e.g., SSO integration). Each application required an access proxy configuration and, in many cases, a specific stanza in the Access Control Engine. Each application went through the following phases:

1. Available directly from the privileged network and via a VPN connection externally.
2. Available directly from the privileged network and via the access proxy from external and unprivileged networks. In this case, we used split DNS. The internal name server pointed directly at the application, and the external name pointed at the access proxy.
3. Available via the access proxy from external, privileged, and unprivileged networks.

Job Function Analysis

By examining job functions throughout the company and cross-referencing this information against the workflow qualification, we were able to prioritize groups of users to migrate. Therefore, we were able to choose network users from the finance, sales, legal, or engineering groups based upon a thorough understanding of user workflows and the capabilities of the BeyondCorp components at that time.

Cutting Back on the Usage of VPN

As more and more applications became available via the access proxy, we started actively discouraging users from using the VPN, employing the following strategy:

1. We restricted VPN access to users with a proven need.
2. We monitored use of the VPN and removed access rights from users who did not use VPN over a well-defined period.
3. We monitored the VPN usage for active VPN users. If all of their workflows were available through the access proxy, we strongly encouraged users to give up their VPN access rights.

Traffic Analysis Pipeline

It was very important that we moved users to the unprivileged network only when we were certain (or very close to certain) that all of their workflows were available from this network. To establish a relative degree of certainty, we built a Traffic Analysis Pipeline. As input to this pipeline, we captured sampled netflow data from every switch in the company. This data was then analyzed against the canonical ACL between the unprivileged network and the rest of the company's network. Such analysis allowed us to identify the total traffic that would have passed the ACL, plus an ordered list of traffic that would not have passed the ACL. The non-passing traffic could then be attached to specific workflows and/or specific users and/or specific devices. We then progressively worked through the list of non-passing traffic to make it function in the BeyondCorp environment.

Unprivileged Network Simulation

To augment the Traffic Analysis Pipeline, which used sampled data from switches, we also simulated unprivileged network behavior across the company via a traffic monitor that was installed on all user devices attached to Google's network. The traffic monitor examined all incoming and outgoing traffic on a per-device basis, validated this traffic against the canonical ACL between the unprivileged network and the rest of the company's network, and logged the traffic that did not pass the validations. The monitor had two modes:

- Logging mode: captured the ineligible traffic, but still permitted said traffic to leave the device.
- Enforcement mode: captured and dropped the ineligible traffic.

Migration Strategy

With the Traffic Analysis Pipeline and the unprivileged simulation in place, we defined and are currently implementing a phased migration strategy that entails the following:

1. Identifying potential sets of candidates by job function and/or workflow and/or location.
2. Operating the simulator in logging mode, identifying users and devices that have >99.9% eligible traffic for a contiguous 30-day period.
3. Activating simulator enforcement mode for users and devices that have >99.99% eligible traffic for that period. If necessary, users can revert the simulator to logging mode.
4. After operating the simulator in enforcement mode successfully for 30 days, recording this fact in the device inventory.
5. Along with inclusion in the candidate set, successful operation in the simulator's enforcement mode for 30 days provides a very strong signal that the device should be assigned to the unprivileged network when the next 802.1x authentication request is serviced by the RADIUS servers.

Exemption Handling

In addition to automating the migration of users and devices from our privileged to our new unprivileged network as much as possible, we also implemented a simple process for users to request temporary exemptions from this migration. We maintained a known list of workflows that were not yet qualified for BeyondCorp. Users could search through these workflows, and with the correct approval levels, mark themselves and their devices as active users of a certain workflow. When the workflow was eventually qualified, its users were notified and were again eligible to be selected for migration.

Completing BeyondCorp

The migration of the Google Enterprise to BeyondCorp is well underway, and the majority of workflows it entails are already qualified. Our migration tools and strategy permit us to proactively move users, devices, and workflows to BeyondCorp without affecting day-to-day productivity.

We anticipate a long tail of workflows that will take some time to move to BeyondCorp. For example, fat-client applications that use proprietary protocols to talk to servers will be a challenge.

We are investigating ways to BeyondCorp such applications, perhaps by pairing them with an authentication service.

As we move forward with the migration to BeyondCorp, we intend to publish subsequent articles explaining why and how Google has moved to BeyondCorp, with the goal of encouraging other enterprises in implementing similar strategies.