PowerMan: An Out-of-Band Management Network for Datacenters Using Power Line Communication

Li Chen, Jiacheng Xia, Bairen Yi, and Kai Chen,
The Hong Kong University of Science and Technology

https://www.usenix.org/conference/nsdi18/presentation/chen-li

This paper is included in the Proceedings of the 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI '18).

April 9–11, 2018 • Renton, WA, USA

ISBN 978-1-939133-01-4

Open access to the Proceedings of the 15th USENIX Symposium on Networked Systems Design and Implementation is sponsored by USENIX.

PowerMan: An Out-of-Band Management Network for Datacenters using Power Line Communication

Li Chen, Jiacheng Xia, Bairen Yi, Kai Chen
SING Group, Hong Kong University of Science and Technology

Abstract

Management tasks in datacenters are usually executed in-band with the data plane applications, making them susceptible to faults and failures in the data plane. In this paper, we introduce power line communication (PLC) to datacenters as an out-of-band management channel. We design PowerMan, a novel datacenter management network that can be readily built into existing datacenter power systems. With commercially available PLC devices, we implement a small 2-layer PowerMan prototype with 12 servers. Using this real testbed, as well as large-scale simulations, we demonstrate the potential of PowerMan as a management network in terms of performance, reliability, and cost.

1 Introduction

A typical datacenter [29, 65, 67] contains more than thousands of servers, switches, storage units, etc. Datacenter operations and management tasks [42, 52, 85] include device installation, bring-up/restart, configuration, monitoring, diagnostics, and Software Defined Networking (SDN) applications [58], etc. At such scale, delivering management traffic is a critical task.

In existing datacenters, management traffic is usually carried in-band with the data plane traffic. Separate service queues and/or VLANs [11] may be reserved for reliable and timely delivery of management messages. However, this approach introduces *fate sharing* [85] between the data plane traffic and management traffic: failures in data plane network will cut off management traffic to the exact network regions at fault, rendering important and relevant management tasks, such as diagnostics and recovery, impossible.

Therefore, an out-of-band management network (MN) is desirable for datacenter operations. A practical out-of-band MN for datacenters should be:

- **Survivable:** MN should be always available, and should survive faults and failures in the datacenter, in order to perform diagnostic and recovery tasks.
- **Scalable:** MN should be scalable enough to access all the devices in the datacenter.
- **Deployable:** MN should be deployable at low cost, and compatible with existing infrastructure.

Prior proposals do not meet these requirements simultaneously. Out-of-band MNs can be constructed as a parallel electrical network[1] using the same networking equipments as the data plane. To reach all devices, this parallel network needs a port count larger than the data plane network; because this fabric not only accesses all the servers like a data plane network, it also needs to reach the management ports of all the switches and other devices. Thus, the cost is prohibitive to build an parallel high port count electric fabric as a MN.

Non-electrical communication channels in datacenters, such as WiFi [36, 47, 84, 85] and free space optics (FSO) [41, 48], are usually built to accommodate dynamic data plane traffic demands. As out-of-band MNs (parallel to data plane network), deploying them results in significant changes to datacenter infrastructure (e.g., raising the ceiling [48, 84], installing reflective surfaces [36, 41, 48, 84], etc). Furthermore, it is also expensive to build a wireless or FSO fabric that reaches the port count required by MNs with current technologies (§6.3).

We believe, for a datacenter MN, power line communication (PLC) technology is an appealing option. PLC [39], proposed in 1900s [66], allows communication between devices connected by power lines. PLC is known to be challenging [60, 69]. However, over short distances and among limited nodes, current PLC modems for home-use can support Gigabit connections using OFDM [61] in PHY layer and CSMA/CA [33] in the MAC layer, providing Ethernet networking to home appliances, e.g., smart TVs, WiFi extender, home networking, etc. We believe these emerging technologies

[1] Although PLC also uses electrical components and electrical wiring to transmit data, for clarity, we use "electrical network" to refer to the electrical packet switching network [19, 43] in the data plane of current datacenters [65, 67].

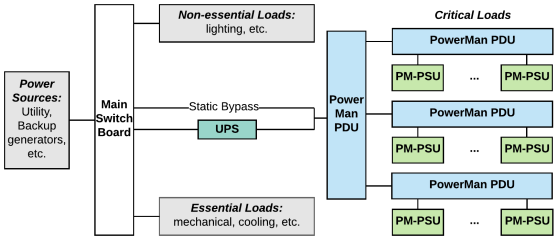


Figure 1: Datacenter Power System with PowerMan

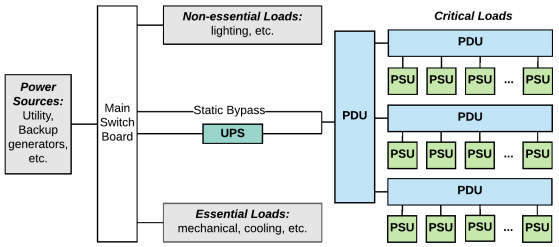


Figure 2: Typical Datacenter Power System (DCPS)

(described by HomePlug standards [24, 56, 81]) open up the opportunity of building a low-cost PLC network with necessary bandwidth and latency for management applications in datacenters.

First of all, a MN using PLC technology naturally meets the survivability requirement, as power system is foundational to any datacenter. A built-in PLC management network is always available, long-term survivable, reaching every device, and independent of the data plane. Furthermore, deployment of PLC network reuses the power system wiring, and requires no change to the existing room designs or floor plan, which is economic.

Yet, one question remains: *can a MN using PLC also meet the other two requirements—being scalable and deployable in existing datacenters?* To answer this question, we build a PLC testbed (§3) with commodity-off-the-shelf (COTS) PLC modems designed for households (using OFDM in PHY layer and CSMA/CA in the MAC layer). Our experiments show that task completion times and user experience of real management applications on PLC network is comparable to that on a Gigabit electrical network. However, we also conclude that a MN directly using COTS PLC devices *cannot* meet the above two requirements: 1) additional in-rack wiring may exceed existing rack designs; 2) due to PLC signal interference, the network can only scale to 6 nodes within a small range (usually 100s of meters [13]) on a single power circuit.

To tackle these problems, we design PowerMan (§4), a MN that can be constructed with existing PLC technology, to support datacenters with more than 10^5 servers. As shown in Figure 1, PowerMan redesigns and replaces two key components in existing datacenter power systems (Figure 2): a power supply unit (PSU) for servers and switches, and a power distribution unit (PDU).

- **PowerMan PSU** lowers the wiring complexity by increasing the integration level. It combines a normal PSU module with a PLC modem module, and acts as a network interface to the server OS. Using PowerMan PSU, PLC network can be deployed easily in the current server racks.
- **PowerMan PDU** addresses the scalability issue. Due to available carrier frequencies and signal quality constraints, COTS PLC modems designed for home-use only support communication within 64 nodes in the same PLC LAN (PLAN) within limited range (§2.2). To scale beyond this, PowerMan PDU eliminates signal interference on the boundary of PLANs using low pass filters, enabling the reuse of the same carrier frequencies across different PLANs. By connecting multiple PLANs into a tree topology, we can scale the PLC network to reach potentially $>10^5$ servers, providing datacenter-wide coverage.

We have implemented a 2-layer PowerMan prototype (§5) connecting 12 servers across two racks. The prototype is built with existing PLC technology in an academic datacenter without any modification to the existing infrastructure, e.g., room plan, power line wiring, and ceiling height. We demonstrate the potential of PowerMan as a datacenter MN (in terms of performance (§6.1), availability (§6.2), and cost (§6.3)) by running real management applications in our small testbed as well as large-scale simulations. Our key contributions are:

- We introduce PLC as an out-of-band channel for datacenter MN. To validate the idea, we build a real small-scale PLC testbed to quantify the throughput, latency, and packet loss conditions for management applications (§3). We find that, due to various sources of interference in datacenter [60], PLC testbed exhibits lower performance than advertised (e.g. $\leq$50Mbps TCP throughput (measured) v.s. 1000Mbps PHY bit-rate (advertised)). We further expose the wiring complexity and scalability issues that cannot be addressed with existing PLC devices.
- We design PowerMan to address the wiring complexity and scalability problems identified above. We validate the design by implementing a PowerMan prototype. On the prototype, our experiments with production traces show $<$24ms average flow completion time (FCT) and $>$10Mbps throughput for the 1-to-N/N-to-1 management traffic patterns. Experiments with real management tasks show that, compared to a Gigabit electrical network, the completion times of all tasks are prolonged by $<$40.62% on PowerMan, with a minimum of 1.57% (66.43s $\rightarrow$ 67.47s) for a Human-in-the-Loop task, and a maximum of 40.62% (32ms $\rightarrow$ 45ms) for a SDN task. We also confirm PowerMan's utility at large scale with simulations, and find that for a PowerMan with 120K servers, the round-trip

time (RTT) for management applications is ∼40ms.

- Cost comparisons with other technologies show that, apart from saving infrastructure modification costs, PowerMan can be constructed with low equipment cost ($1/2$∼$1/3$ of the cost of related designs at the same scale). PowerMan is also power-efficient in operation: its power consumption is 6%∼9% of other technologies.

Caveat: PowerMan is suitable for many management applications given its performance characteristics. However, we acknowledge that, for some applications, delivering control messages with low latency is crucial: fine-grained load balancing [62, 82] and flow scheduling [20, 30] need to configure data plane on millisecond time scale. *PowerMan alone is not suitable for such applications.* For them, we suggest dual-homing the controller with access to both the data plane network and PowerMan network. Latency-sensitive traffic can use the fast data plane, while PowerMan can serve other management applications. We believe PowerMan is also valuable as a back-up/diagnostic network to fall-back on in case of failures.

2 Background

2.1 Power System in Datacenters

The power system [35, 44] is the most fundamental system in a datacenter. A typical DCPS is shown in Figure 2, and it is composed of:

- The main switch board (MSB) directs electricity from one or more sources of supply to several smaller regions of usage. It feeds into all loads in the datacenter.
- The uninterruptible power supply (UPS) provides consistent power to critical loads without interruption. It contains energy storage, which supplies power to the load when the utility power is down.
- A power distribution unit (PDU) is an electrical distribution device, and it can be free-standing or rack-mounted. The PDU houses circuit breakers that are used to create multiple branch circuits from a single feeder circuit, and can also contain transformers, surge protection devices, and power monitoring/controls.
- A power supply unit (PSU) rectifies AC power from the connected PDU to DC power. For reliability, critical servers and switches are usually equipped with two PSUs in case of failure.

DCPS is often classified as belonging to "Tier I-IV" [71] depending on the power distribution, UPS, redundancy, etc. [15, 18] For example, in a Tier-III DCPS [35, 44], each critical load device has two power distribution paths (including redundancy components), and the power system in Figure 2 is replicated for each PSU. In what follows, for clarity, our design of PLC networking is limited to the primary power system depicted in Figure 2 by default, and we discuss how all tiers of DCPS can adopt PowerMan in §4.4.

2.2 Power Line Communication

PLC uses electrical wires to simultaneously carry high frequency data signals and 50∼60Hz AC power transmission. PLC has been widely used in power systems for protection, telemetering, and industrial control applications [39, 40, 45, 83] with data rate of a few Kbps.

In recent years, we witness a rapid growth of PLC appliances for home networking, due to its ubiquity (home power systems provide sockets in every room) and ease of deployment (no new wire needed). The home networking market drives PLC technology to reach higher bandwidth, in order to support popular use cases such as broadband Internet access, video streaming, gaming, etc. Through standardization efforts from the US HomePlug Powerline Alliance and European Home System Consortium, vendors have converged to use Orthogonal Frequency Division Multiplexing (OFDM) [61] as the modulation scheme in PHY layer, and CSMA/CA [33] as the MAC layer protocol. Adopting the HomePlug protocols (HomePlug 1.0 [56]/AV [24]/AV2 [81]), PLC modems and adapters can form a communication network providing Ethernet connectivity to TVs, gaming console, and PCs. Currently, many vendors offer PLC modems with up to 1200Mbps PHY layer bit-rate [8, 16, 17], and up to 64 devices in one PLC network [13, 16, 17].

In academia, there have been continuous efforts in PHY [60, 69] and MAC [74, 75, 76, 77, 78] layers for PLC to achieve higher throughput and lower latency. Orthogonal to prior work, we focus on the application of PLC in the context of datacenter MN. We believe PLC is a suitable candidate for MN as a built-in communication network in DCPS for the following reasons:

- Power system is the last-to-fail system in datacenters, and is independent of the data plane network. A MN within the power system therefore can survive data plane failures, and is ready for immediate diagnosis and recovery.
- Power system reaches every device in datacenters, providing full visibility for management applications.
- PLC reuses the wires in the power system, and there is no need to change the existing room plans, ceiling height, and rack dimensions. This compatibility with existing datacenter designs greatly reduces the deployment cost.

In the following, leveraging the technology advances in household PLC appliances, we are motivated to: 1) understand performance characteristics of PLC networks (§3), 2) expand PLC network from home-scale to datacenter-scale (§4), and 3) evaluate our design for datacenter MN using PLC (§6).

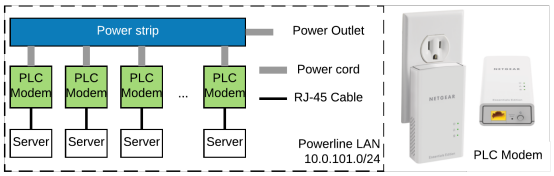


Figure 3: COTS PLC Testbed & PLC Modem

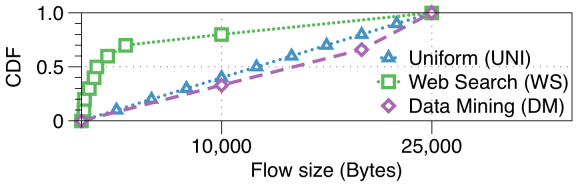


Figure 4: Flow Size Distributions

3 Building a PLC Testbed

In this section, we first build a small-scale PLC testbed. Then, we perform a series of experiments and measurements to learn the communication characteristics of the PLC channel in a datacenter environment.

3.1 Building a PLC Testbed

We describe the devices we use and how they are connected to build the testbed.

Server & switch: We use Huawei FusionServer RH1288 with Intel E5-2630 and 64GB memory (1 Rack Unit). Each server has a NetXtreme BCM5719 Ethernet Network interface card (NIC) with 4×1 GbE ports. The servers are all connected to a Gigabit Ethernet switch via their first Ethernet interface (Eth0).

PLC modem: We obtained 16 Netgear Powerline 1000 (PL1000) PLC modems (US$ 30.3 per piece) via local home appliance vendors. As in Figure 3, each modem has one built-in power plug and one RJ-45 port for Ethernet connection. The max power consumption of PL1000 is 3.73 watts (0.49 watts in standby mode). It is compatible with HomePlug AV protocols. For OFDM, it uses frequencies in the range from 2 MHz to 86 MHz.

PDU: The rack-mounted servers and Ethernet switch are plugged into the in-rack PDU with no empty sockets. We use a separate Thomson TM-EC6 8-socket power extension cord for PLC modems.

Interconnection & wiring: We connect the PLC modems to the power extension cord, and then plug them into the power outlets on the in-rack PDU. Each server is connected to one PLC modem via its second Ethernet interface (Eth1).

In summary, as shown in Figure 3, we build a PLC testbed using commodity components. Each server is both connected to an electrical ToR Ethernet switch via Eth0, and to a PLC modem via Eth1. These modems are connected via a power strip, forming a PLC network.

Through building the testbed, we identify the first difficulty for practical deployment of PLC networking in datacenters: wiring. This is because each of these external PLC modems requires an additional power socket slot and a network cable, resulting in 2× socket count on the in-rack PDU and 1.5× space for cabling. As the current rack design does not anticipate the usage of PLC devices, we find it difficult to organize the additional cables, and the PLC modems have to be attached to a power extension cord. We will address this in §4.1.

Pattern	CDF	AFCT-us	99% FCT-us	Thruput-Mbps	Pkt Loss%
1-to-1	DM	3887 (233)	8631 (327)	48.15 (484.56)	0.00% (0.00%)
1-to-5	DM	12914 (686)	29552 (1146)	35.39 (791.49)	0.13% (0.00%)
5-to-1	DM	16429 (606)	43210 (944)	33.52 (931.06)	0.12% (0.00%)
1-to-1	UNI	3972 (223)	8686 (331)	25.00 (444.11)	0.00% (0.00%)
1-to-5	UNI	11590 (618)	26798 (1143)	30.48 (763.02)	0.13% (0.00%)
5-to-1	UNI	15728 (532)	39639 (870)	31.13 (928.74)	0.13% (0.00%)
1-to-1	WS	2895 (187)	7900 (321)	13.11 (202.57)	0.00% (0.00%)
1-to-5	WS	9234 (337)	31049 (1117)	13.98 (522.98)	0.23% (0.00%)
5-to-1	WS	11021 (296)	36435 (618)	17.00 (635.87)	0.17% (0.00%)

Table 1: Measurements of Synthetic Traffic on PLC Testbed. The results of a gigabit electrical network are shown in the parentheses.

3.2 Testbed Experiments

Next, we measure its performance using both synthetic traffic and real management applications.

3.2.1 Scalability

We first investigate how many PLC modems can coexist in a PLC network. We add PLC modems to the power strip one by one (IP addresses and subnet masks are assigned beforehand), and then monitor the indicator lights on the modems for successful connections. Finally, we verify the connection on the servers via `ping` utility. We observe that the network can accommodate at most 6 PLC modems. When there are more than 6 modems in the network, the first 6 modems are connected.

3.2.2 Experiments with Production Traces

Setting: For the flow size, we adopt 2 realistic flow size distributions used in prior work [21, 22, 25, 43]: one from a web search cluster [21] and the other from a data mining cluster [43], respectively. We also include a uniform distribution for reference. All distributions, shown in Figure 4 are capped at 25KB, as we are mainly interested in management applications, which tend to have shorter flow sizes.

We use the following traffic patterns:

- **1-to-N:** This pattern occurs in management applications where a master pushes configurations to slaves.
- **N-to-1:** This pattern occurs in monitoring applications where a server collects statistics from clients.

Among the 6 connected servers, we create traffic patterns using a traffic generator [6], which is a client/server application for generating user-defined traffic. The server listens for incoming requests on the specified ports, and replies with a flow with the requested size for each request. The client connects to a list of servers, and generates requests to randomly chosen servers. For each

request, it samples from the input request size and fanout distributions to determine the request size and how many flows to generate in parallel for the request. All packets use the same default priority.

In each experiment, we use a different combination of patterns and distributions, and each client generates 25K requests. We measure the flow completion time (FCT), throughput, and packet loss rate for each flow, and Table 1 summarizes the results. The average round-trip time (RTT, grouped by traffic patterns) is shown in Figure 5. We then repeat the experiments using the Gigabit electrical network (via Eth0), and the results are shown in the parentheses in Table 1.

Results: We make the following observations:

- **Latency:** The average FCT on PLC testbed is around 2 order-of-magnitude (OoM) larger than that on the electrical network. We observe the same trend for 99th percentile FCT. For RTT, the smallest one (2.2ms, from 1-to-1 pattern) is also around 2 OoM larger (∼20us on the electrical network).
- **Throughput:**The advertised 1000Mbps bandwidth is in fact the maximum PHY bit-rate, and we cannot obtain more than 50Mbps TCP throughput on the testbed, which matches field-tested results [23]. The throughput is about 1 OoM less than that on the electrical network.
- **Packet loss rate:** The packet loss rates are less than 0.5% for PLC testbed across all cases, while the electrical network shows near-zero packet loss rates.

Implications: As expected, the PLC testbed we constructed shows much lower throughput and longer latency compared to the Gigabit electrical network. This is because PLC is an "extremely harsh environment" [60] for the high-bandwidth, high-frequency communication signals, as critical channel parameters (e.g., noise, impedance, and attenuation) are highly unpredictable and varied with time, frequency and location [69]. As a result, the PLC network is clearly inappropriate for time-critical tasks (e.g., fine-grained load balancing [62, 82] and flow scheduling [20, 30]).

However, the PLC testbed is shown to deliver <10ms average FCT and >10Mbps throughput for N-to-1/1-to-N patterns, which are common for management applications. Thus, for latency-insensitive management tasks (e.g., device installation, bring-up/restart, configuration, monitoring, diagnostics, etc), the PLC network remains attractive, due to its other benefits such as survivability in case of data plane failure, compatibility with existing datacenter design, and economy.

Therefore, we proceed to evaluate the end-to-end application performance of the PLC testbed with latency-insensitive management tasks.

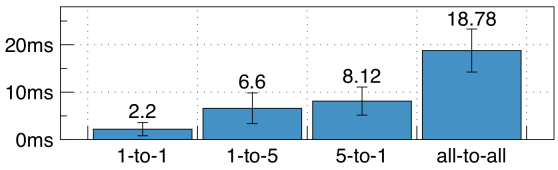


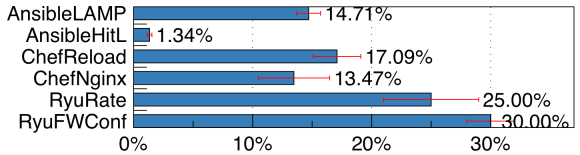
Figure 5: RTT measurements on PLC testbed

3.2.3 Experiments with Management Applications

Setting: Based on the communication model, we choose two management platforms for on-premise datacenters and cloud virtual clusters: push-based Ansible [1]), and pull-based Chef [3]. As SDN is an important class of applications, we also include a SDN controller, Ryu [12]. We perform tasks with real usage scenarios.

- **Ansible [1]** is an automation engine for clusters. Ansible is push-based and agentless: from the master node, it manages slave nodes through SSH connections. We deploy Ansible 1.7.2 on our testbed, and perform one automated task and one Human-in-the-Loop (HitL) task:
  - **AnsibleLAMP**: An automated LAMP deployment with two web severs, two load balancers, and two database servers. The playbook is based on Ansible official examples [2].
  - **AnsibleHitL**: A HitL setting with an operator checking configurations of servers. Via Ansible ad-hoc commands [7], in each experiment, the operator sequentially executes `df`, `route`, and `lsmod` on all servers.
- **Chef [3]** is an automation platform for cluster management. Chef is a pull-based: clients poll a centralized master periodically for updates. On our testbed, we install a Chef Server 12.11 in standalone mode on one of the servers, and the rest are installed with Chef Client 12.17. We perform two automated tasks described by Chef cookbooks.
  - **ChefReload**: This cookbook [4] automatically reloads the Apache service on all servers.
  - **ChefNginx**: This cookbook [10] automatically distributes the install file (889KB), installs, and configures nginx [9] 1.10.2 on all servers.
- **Ryu [12]** is a SDN framework. It can be integrated with OpenStack Neutron [5] for SDN applications. We installed OpenVSwitch 2.5.1 [63] on all servers and a Ryu 3.26 controller on one of them. We run two tasks in official documentation [14]:
  - **RyuRate**: We use `curl` to query the Ryu controller via its RESTful API, and the controller replies with the current rates of all ports.
  - **RyuFWConf**: We add a firewall rule via RESTful API, and the Ryu controller replies with the result.

We run the above 6 management tasks (each for 10 times) and measure their completion times with millisecond precision on both the PLC network and Gigabit



	AnsibleLAMP	AnsibleHitL	ChefReload	ChefNginx	RyuRate	RyuFWConf
Elec. Network	273.45s	66.43s	17.91s	14.77s	0.032s	0.040s
PLC Testbed	313.70s	67.32s	20.97s	16.76s	0.040s	0.052s

Figure 6: Management Applications on PLC testbed

electrical network. We use percentage increase in completion time as the metric: for a task, denote its completion time on Gigabit electrical network as T_e and on PLC testbed T_p, the metric is defined as: $\frac{T_p - T_e}{T_e} \times 100\%$.

Results: In Figure 6, we observe encouraging application performance delivered by the PLC network. Overall, we find that using PLC results in less than 30% increase in completion time compared to the Gigabit electrical network for all the tasks, and this increase is mainly due to the latency introduced by the PLC network. In the best case, we notice that, for AnsibleHitL task, the PLC network performs almost the same as the electrical network (only 1.34% longer). This is because human response time is the main contributor of latency in this task. In the worst case, the completion time of RyuFWConf is increased by 30%, which is because it performs only HTTP query/response and network latency contributes the most to the completion time. In summary, our results of end-to-end application performance on PLC network is promising for latency-insensitive management tasks.

3.3 Lessons Learnt

We conclude: 1) It is possible to use commodity PLC modems to form a PLC network that provides Ethernet connectivity for all connected servers. 2) PLC performance is promising for management applications: it provides <10ms average FCT and >10Mbps throughput for N-to-1/1-to-N patterns, which are common for management applications; the management tasks also have similar user experience. 3) This PLC network, however, cannot be directly used in real datacenters due to the deployability (wiring) and scalability problems.

Therefore, we are motivated to tackle the wiring and scalability issues, so that the PLC technology can be deployed in real datacenters.

4 PowerMan Design

To tackle the wiring and scalability issues, we design PowerMan. To ensure deployability, our guiding principle is to respect the existing datacenter designs, and preserve the floor plan, room design, rack dimensions, and power line wiring. To this end, PowerMan only replaces two types of components in existing DCPS: PSU and PDU.

4.1 Power Supply Unit (PSU)

In the PLC testbed, each server needs two network cables: one for data plane connectivity, and one for PLC

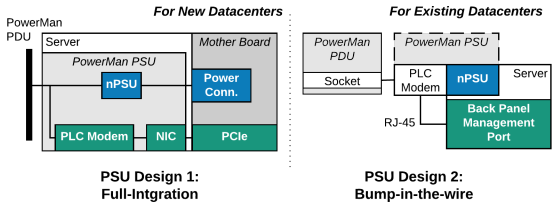


Figure 7: PowerMan PSU

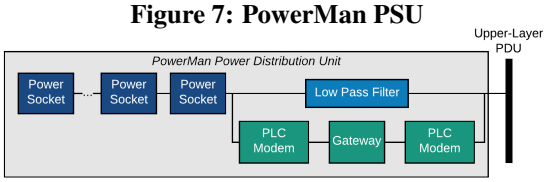


Figure 8: PowerMan PDU

modem to access MN. It also needs two power sockets: one for its PSU, and one for the PLC modem. Thus, a PLC MN requires 1.5× cable space and 2× power sockets per rack. Given a rack hosts 20∼40 servers [65, 67], it is infeasible to accommodate this additional wiring with existing rack design. To address the problem, we design a novel PSU for rack-mounted servers and switches in datacenters (Figure 7). The key idea is to increase the integration level in the PSU to reduce external wiring.

We present two designs of PowerMan PSU to fit different deployment scenarios: Full-Integration and Bump-in-the-Wire. Full-Integration is designed for new installation of datacenters, as the datacenter operator has the freedom to customize the hardware configuration of each server/switch. We combine a normal PSU module (nPSU in Figure 7), a PLC modem module, and a network interface module in the PowerMan PSU. It connects to the mainboard of the server via a PCIe interface, and appears as another NIC to the OS, which allows users to use familiar networking stack to access the PLC network.

The Bump-in-the-Wire design is for incremental deployment in existing datacenters, and it leverages the integrated NIC on the mainboard of rack-mounted servers, which is exposed as the management port on the server back panel. The PLC modem attaches to the PSU externally, and acts as a "bump" in the power cable from the PDU socket to the PSU. The power to the server is fed into its PSU through the PLC modem via a bypass circuit. The PLC modem connects to the management port via a RJ-45 network cable, so that the integrated NIC can access the PLC network. This network cable travels a short distance from the power port to the management port on the back panel, and thus does not tangle with other in-rack cables.

Via PowerMan PSU, a server can connect to a PLC MN without complicated wiring and additional power sockets, thus is compatible with the design and dimensions of the current racks.

4.2 Power Distribution Unit (PDU)

The scale of PLC network on our testbed is limited to 6 nodes. We refer the PLC network within the PDU as PLC LAN (PLAN). Manufacturers of more advanced models claim that the scale can be as large as 64 nodes [13, 16, 17]. However, this is still too small for production datacenters [65, 67]. The main reason for such limited scalability is that these devices are designed for home-use, where network size is not the main concern.

To scale, we design a novel rack PDU (Figure 8). The key idea is to remove cross-PDU PLC signal interference but maintains network connectivity. PowerMan PDU achieves this with two main components, a low pass filter (LPF) and a PLC gateway.

We keep the circuit of a normal PDU, and add a LPF between the circuit and the external power line. Since the OFDM frequencies used in the PLC modem is $\geq$ 1.8MHz [60] and the AC power frequency is 50$\sim$60Hz, a LPF with appropriate cut-off frequency (between 60Hz to 1.8MHz) can greatly attenuate the outgoing and incoming high frequency PLC signals, thus effectively eliminating the interference from/to other PLANs.

While the PLC signals are mostly eliminated across the LPF, the network connectivity is preserved using a PLC gateway. The PLC gateway consists of a packet-forwarding hardware gateway and two PLC modems. One modem is connected to the PLC network inside the PDU, and the other is connected to the PLC network on the external, upper-layer PDU. The PLC gateway is therefore connecting the PDU's PLAN and the upper-layer PDU's PLAN by forwarding packets between them, with no PLC signal interference.

PowerMan PDU replaces the rack PDU and retains the same cable and socket count. It acts as a switch for the PLC network devices on the same rack.

4.3 Interconnection & Scalability

With the new PowerMan PDUs developed, we can now connect them and scale the PLC network to support real datacenters. We leverage DCPS to interconnect the PLC devices. Since PDUs in DCPS are connected in a tree topology (Figure 2), we also choose to use the same topology to scale. Other topologies (e.g. ring, mesh, hypercube, etc.) requires changing the wiring of the power system. Take ring topology as an example, each PDU connects to more than one other PDUs, requiring an additional power cable for each PDU. Other topologies also requires a different power allocation scheme, both inside the PDU and across PDUs.

As shown in Figure 9, we construct the PowerMan PDUs into a $(k-1)$-ary tree topology, where k is the number of PLC devices supported in a PLAN. For our current PLC modems, k=6; up to k=64 have been reported for other COTS PLC modem models [13]. With height

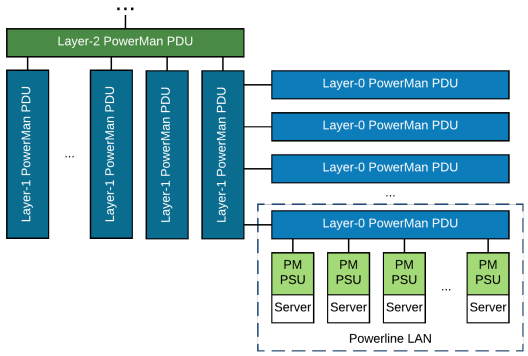


Figure 9: Scaling PLC with PowerMan

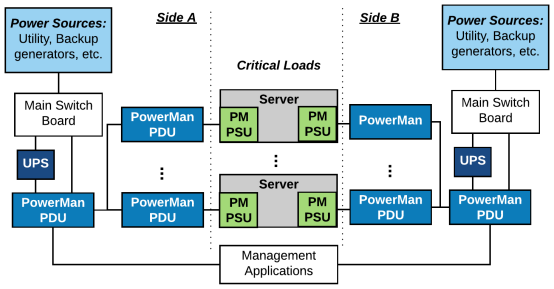


Figure 10: PowerMan Fault-Tolerance: Example of Tier-III DCPS with AB Dual-Bus [44]

h, this topology can connect $(k-1)^h$ PowerMan PSUs. With a tree height of h=3 and k=64, 250K PSUs can be connected.

4.4 Fault-tolerance

PowerMan leverages the redundancy in existing DCPS to achieve high availability. As mentioned in §2, DCPS can be classified into 4 tiers [71], and all can be integrated with PowerMan.

- **Tier-I** DCPS have a single path for power distribution without redundant components, and PowerMan can be integrated as in Figure 9.
- **Tier-II** adds redundant components to this design (N + 1), improving availability, and PowerMan can be integrated into the main distribution path as for Tier-I DCPS, the PDUs in the redundant components should also be replaced with PowerMan PDUs.
- **Tier-III** datacenters have one active and one alternate distribution path for utilities. Each path has redundant components and are concurrently maintainable, providing redundancy during maintenance. PowerMan can be integrated into both distribution paths. As an example, Figure 10 showcases how PowerMan can be integrated with Tier-III DCPS [44]. This architecture is configured with two sides, A and B. Each side can include multiple UPSs, and either side can handle 100% load. If one side has a problem, the load

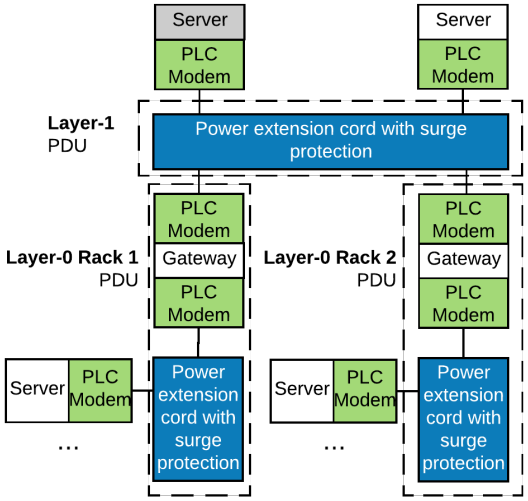


Figure 11: Two-Layer PowerMan Prototype

is automatically switched to the other side. Automatic load transfer switches can reside upstream of the UPS for maintenance isolation purpose. This design ensures a high level of system availability even during maintenance or component failure. PowerMan can therefore be replicated in both sides; the controller node where management applications are located should also connect to the root PowerMan PDUs on both sides, so that when failure (either PowerMan or DCPS) in either side happens, management applications can still access the servers and switches.

- **Tier-IV** DCPS have two simultaneously active power distribution paths, redundant components in each path, and are supposed to tolerate any single equipment failure without impacting the load. PowerMan can be integrated in the same way as Tier-III.

Embedded in DCPS, PowerMan share the redundancy and availability mechanisms, thus is expected survive even partial power outages in Tier-II (or higher) DCPS.

5 Prototype Implementation

We implement a PowerMan prototype to validate the design, and its schematic is shown in Figure 11. We have not yet constructed a PowerMan PSU that can be fit into our rack-mounted servers, but its functionality can be emulated using the same setting as §3.1: each server connects to a PLC modem via one of its NIC ports (Eth1).

PDU has two components: a LPF and a PLC gateway. For the LPF, instead of implementing LPF circuits and installing them on the power extension cords, we identify that power extension cords with surge protection can serve as low cost alternatives[2]. This is because surge protector removes voltage spikes and high frequency noise. We note that the use of surge protector as LPF is only for prototyping, and real deployment of PowerMan should use properly designed LPF in the PDU. We obtained Targus SmartSurge 6 power extension cords from local vendors, and our testing shows that two PLC modems cannot establish connection across two such cords, which indicates that they have the correct cut-off frequency. In this way, the PLC modems can form a PLAN within the power extension cord that they are attached to, without interference of PLC signals from other PLANs.

[2]This choice is inspired by the product FAQ [13] from the vendor of our PLC modem. The FAQ advices against the usage of surge protectors with the PLC modems, because surge protector may remove high frequency signals.

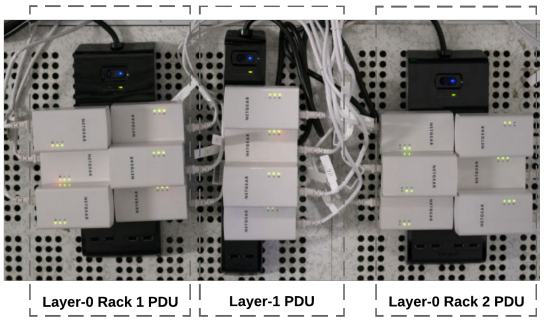


Figure 12: PLC Network Components of PowerMan prototype

Next, we implement the PLC gateway with two PLC modems and a rack-mounted server. The server connects to the two modems via Eth1 and Eth2 ports. The modems are attached to different power extension cords with surge protection. Therefore, their signals are isolated, and can only propagate within their own PLANs. With routing rules correctly configured, the server acts as a packet forwarding gateway between the two PLANs.

We construct 3 prototype PowerMan PDUs, which form a tree topology with 2 layers, as shown in Figure 11&12. In Layer-0, the prototype has two racks, and each rack forms a PLAN on its own PDU. The two Layer-0 racks are connected to a Layer-1 PLAN via their PLC gateways. In addition to the gateways of the two racks, we connect another two servers to act as gateways on the Layer-1 PLAN. The routing tables and IP addresses are properly configured in all the servers and gateways, so that each server can reach all the other servers on this PowerMan PLC network.

6 Evaluation

In this section, we evaluate three aspects of PowerMan: performance, reliability, and cost.

Summary of results:

- Experiments with production traces show <24ms average FCT and >10Mbps throughput for 1-to-N/N-to-1 traffic patterns.

- Experiments with real management applications demonstrate that, compared to the Gigabit Etherent, the completion times of all tasks are only prolonged by <40.62% on PowerMan.
- By simulating a year of operation, PowerMan is shown to achieve >99.9977% availability (leveraging the redundancy in DCPS) at the scale of 250K servers.
- Apart from saving infrastructure modification costs, PowerMan can be constructed with low initial cost (1/2∼1/3 of the cost of other technologies at the same scale), with 6%∼9% operating power usage.

6.1 Performance

6.1.1 Prototype Experiments

On the PowerMan prototype, we perform the same set of experiments as in §3.2.

Experiments with Production Traces: In addition to the setting in §3.2.2, our experiments here include another parameter: distance, which refers to the number of PLC gateways (i.e., hops) between the servers and clients. For example, for 1-to-5 pattern with distance=1, a client in Rack 1 will only send requests to the traffic generator server hosted in gateways in Layer-1 PLAN[3]. To understand this parameter in real PowerMan deployments, for a controller node connected to the root with tree height h=3, its distance to all PSUs is merely 2. We summarize the results from the experiments on the prototype in Table 13. We make the following observations.

- **Latency:** Compared to Table 1, we see on average 3.04ms increase in FCT if distance increases by 1, and 4.13ms if distance increases by 2. This corresponds to our RTT measurements on the prototype in Figure 14: when distance increases from 0 to 1, the RTT increases on average 2.19ms, and 2.92ms from 1 to 2.
- **Throughput:** Increasing distance by 1 (2) decreases the throughput by 3.27Mbps (6.80Mbps) on average. Still, the prototype provides >10Mbps for 1-to-N/N-to-1 patterns.
- **Packet loss:** interestingly, increasing distance lowers the packet loss rate: 1 (2) increase in distance decreases the packet loss rate by 0.05% (0.05%) on average. This is because the inter-PLAN flows converge at the gateway, and from there, are forwarded to their destinations. This store-and-forward behavior for flows across PLAN results in lower packet loss rate compared to the flows within a PLAN running CSMA/CA.

In summary, PowerMan prototype demonstrates <24ms average FCT and >10Mbps throughput for common management application traffic patterns (1-to-N/N-to-1) for distance=2. This indicates that, a PowerMan with tree height h=3 can support management applications with reasonable latency[4] and throughput.

Pattern	CDF	Distance	AFCT (us)	99% FCT (us)	Thruput (Mbps)	Pkt Loss%
1-to-1	DM	1	7963	15671	32.54	0.00%
1-to-1	DM	2	13856	26245	31.35	0.01%
1-to-5	DM	1	14736	30747	27.52	0.04%
1-to-5	DM	2	19701	39326	24.55	0.03%
5-to-1	DM	1	17418	38063	31.93	0.04%
5-to-1	DM	2	23046	48150	23.89	0.02%
1-to-1	UNI	1	7529	16575	13.19	0.01%
1-to-1	UNI	2	11841	24255	8.39	0.01%
1-to-5	UNI	1	14231	31086	26.22	0.06%
1-to-5	UNI	2	19715	41289	20.46	0.03%
5-to-1	UNI	1	17148	40939	28.29	0.05%
5-to-1	UNI	2	21833	47630	22.25	0.03%
1-to-1	WS	1	5825	15648	6.52	0.02%
1-to-1	WS	2	9601	25792	3.96	0.05%
1-to-5	WS	1	11574	33556	12.19	0.14%
1-to-5	WS	2	14657	38995	10.56	0.07%
5-to-1	WS	1	11783	33270	15.62	0.06%
5-to-1	WS	2	15561	40009	12.13	0.04%

Figure 13: Measurements of trace-based experiments on PowerMan prototype

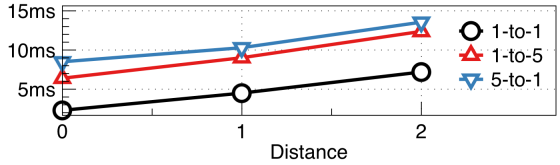
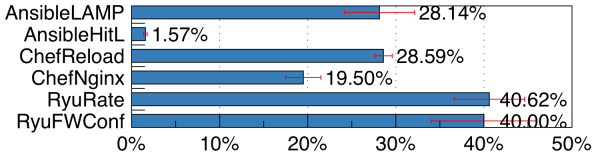


Figure 14: RTT on PowerMan prototype



	AnsibleLAMP	AnsibleHitL	ChefReload	ChefNginx	RyuRate	RyuFWConf
Elec. Network	273.45s	66.43s	17.91s	14.77s	0.032s	0.040s
PowerMan	350.39s	67.47s	23.03s	17.65s	0.045s	0.056s

Figure 15: Management Applications on PowerMan prototype

Experiments with Real Management Applications: Next, we evaluate the end-to-end applications performance. We perform the tasks in §3.2.2 again on both PowerMan prototype and the Gigabit electrical network. We scale the set of tasks in §3.2.3 so that they can cover all 10 servers in the testbed. For example, the AnsibleLAMP task now configures 4 web servers, 2 load balancers, and 4 database servers. We assign one of the gateway server in Layer-1 PLAN as the master node for Ansible, Chef, and Ryu, which is the darkened gateway in Figure 11. We plot the results in Figure 15.

As expected, due to the need of traversing one PLC gateway, the completion times increase for all the tasks. Among them, for AnsibleHitL, the PLC network performs almost the same with the electrical network (only 1.57% slower) as distance increases.Also, as explained in §3.2.2, network latency dominates the completion times of the two Ryu tasks, so their metrics increase the most, i.e., 40.62% and 40% respectively. Furthermore, using PowerMan results in <30% increase in comple-

[3]The setting for the results in Table 1 can be considered as distance=0 (within the same PLAN).

[4]We consider soft real-time constraints for interactive systems, e.g. 300ms [28, 72], are reasonable latency targets.

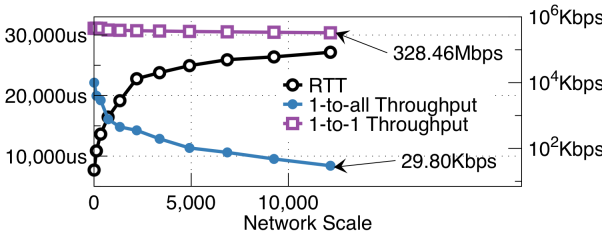


Figure 16: Large Scale Simulations: RTT and throughput

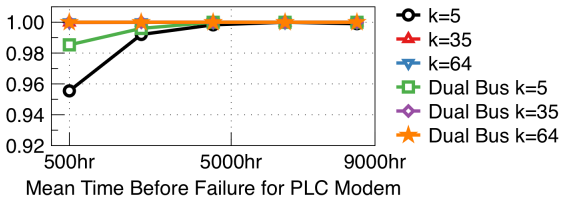


Figure 17: Availability of PowerMan

tion times for Chef tasks. Overall, adding one more tier in PowerMan prototype compared to §3's testbed results in less than 5% increase in task completion time.

6.1.2 Large-scale Simulations

The current prototype is still too small to reveal PowerMan's performance in actual deployments. Using ns-3 [50] simulator, we perform simulations at the scale of real datacenters [65, 67] to infer the user experience of PowerMan in actual deployments.

Setting: Since each PowerMan PDU corresponds to a PLAN that uses CSMA/CA (§2.2), we simulate PLAN using the CSMA network implementation in ns-3 with parameters in [73, 79]. We interconnect PLANs with point-to-point links, which corresponds to the PLC gateways in our design. We assume that a controller connects to the root of the PowerMan tree topology with a 1Gbps network interface. We fix the tree height h=3, so the distance from the controller to every PSU is 2. We first tune the parameters to fit the results in Figure 5&14, so that the RTT within a PLAN is 8ms and the latency across a point-to-point link is 3ms. Then we run the simulations for different scale of the network (number of servers) from 125 (k=6) to 12167 (k=24).

Results: We create 1-to-1 and 1-to-N patterns from the controller using TCP connections, and measure the RTT and throughput per-server for different network scales. The results are plotted in Figure 16. For 1-to-1 connection from the controller to a server, we observe consistent throughput >328.64Mbps. For 1-to-N pattern, we create connections from the controller to all servers, and see that the per server throughput quickly drops as more and more servers shares the out-going bandwidth of the controller. At maximum network scale (12167), the per-server throughput is 29.8Kbps. For latency, we observe that the average RTT is smaller than 40ms even when network scales beyond 10^5 servers. This is as expected as the overall distance is only 2.

6.2 Availability

We use availability to characterize the system reliability of PowerMan, which is the percentage of reachable servers. The key component in PowerMan PDU and PSU is the PLC modem, so we model availability of the entire system at the resolution of an PLC modem. We use a Poisson process [27] to characterize the failure process of a PLC modem. The Mean Time Before Failure (MTBF) is the key metric in this model, and it is a common measure of reliability of a hardware component [27, 53]. A higher MTBF means the component is more reliable. Our PLC testbed and prototype have been running for 2 months without failure (1440 hours); using this as reference, we vary the MTBF of PLC modem from 500 to 9000 hours. The MBTF of packet forwarding gateway is assumed to be 3000 hours [37].

We implement an event-driven failure simulation, modeling the entire network of PLC modems in PowerMan. In each run, we vary the scale of network (by increasing k from 5 to 64, h=3), the MTBF of PLC modem (from 1000 to 9000), and simulate a year of PowerMan operation with the failure model describe above. We plot the average availability of PowerMan in Figure 17. We observe that PowerMan is highly available at large scale: For k=64 (network scale is 250K), the availability is 99.9943% (using the least reliable modem with MBTF=500hrs). High availability provides consistent global visibility to management applications, allowing them to perform monitoring and diagnostic tasks.

In Figure 17, we also plot the availability of a PowerMan in a DCPS with Dual Bus redundancy as shown in Figure 10. In this setting, we have PLC networks replicated in both sides, and the controller attaches to the roots of both trees. We can see that, by integrating with the redundant power systems, PowerMan can achieve higher availability for varying network scales.

Since PowerMan is embedded in DCPS, servicing/replacing components is similar to that in a typical DCPS. Tier II-IV DCPSs are designed with redundancy (§4.4), so when parts of the system fail, the operations can continue, as back-up units will take over. In the meantime, faulty components can be repaired/replaced. PowerMan adopts the same recovery strategy.

6.3 Cost Comparisons

Next we compare the construction, equipment, and operational costs of PowerMan and other related designs that can be used as out-of-band MNs. The comparison is done at the same scale of 16000 servers. We compare with these proposals for datacenters: 3D-Beamforming (3DBF) [84], Firefly [48], Diamond [36], and Fat-Tree [19]. We emphasize that this is not a direct comparison: these designs are complete datacenter networks with both data plane and in-band control plane, and we

Components	FatTree	3D-Beamforming	Firefly	Diamond	PowerMan
NIC (k$)	80	80	80	240	80
Switch (k$)	2080	2080	416	832	0
Wireless (k$)	0	192	2400	1920	0
Cable (k$)	80	80	0	32	0
PLC Modem (k$)	0	0	0	0	787
Gateway (k$)	0	0	0	0	351
Total (k$)	2240	2432	2896	3024	1218

Table 2: Comparison of Equipment Costs

use them as proxy for comparing different technology that can be used to construct out-of-band MNs: 60GHz WiFi, FSO, and electrical packet switching.

Construction cost: Wireless designs (3DBF, Firefly, & Diamond; as well as other WiFi and FSO designs [41, 47]) have various requirements on the datacenter interior designs. For example, reflective surfaces (static [36, 48, 84] or mechanically controlled [41]) must be installed for connectivity. In addition,3DBF has ceiling height requirements [84], which may incur room modifications in deployment. Furthermore, Diamond also requires the spacing between racks. This limits the number of racks per room, and Diamond deployments may need more rooms to hold the same number of servers.

In contrast, PowerMan leverages the wiring in existing DCPS to achieve scalable connectivity for MN, and only replaces the PDUs and PSUs in the DCPS. Thus, constructing PowerMan should incur no cost in modifications of room design or floor plan, which greatly reduces the cost of deployment compared to the other proposals.

Equipment cost: Next, we compare the equipment costs in Table 2, and we explain the assumptions as follows. For PowerMan, we assume PSU uses Design 1, which incurs no cable cost. The tree topology of PowerMan is configured as h=3,k=27. Each PLC modem is \$30[5]. Each Gateway is \$500. For other designs, we consider the cost of NICs on the server, switches, wireless radios and cables. We adopt the conservative estimates in [36], and make the following cost assumptions: each wireless radio component costs \$60 [85], each 40-port switch costs \$1040, each NIC port costs \$5 [46], each FSO device port costs \$150 [48], and an average cost of \$1 per meter for cabling [48] and \$1 per square meter of absorbing paper. We assume the reflectors used [36, 48, 84] have negligible cost as equipments.

Overall, PowerMan can be constructed with $1/2\sim1/3$ of the cost of other proposals at the same scale, confirming PowerMan as a cost-effective option for MN.

Power consumption: Power consumption is an important component of operational cost. In Figure 18, we compare the operational power consumption of different designs. We assume each NIC consumes 5 Watts (W) [36], each PLC modem 3.73W (§3.1), each switch 170W [36], and each gateway 300W[6]. For wireless designs, each FSO component in Firefly consumes 3W [48] and the WiFi module in Diamond 60W [36].

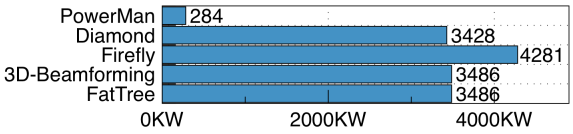


Figure 18: Comparison of Power Consumption

In general, PowerMan consumes much lower power at the same scale, using $6\%\sim9\%$ power of other designs. This is because PowerMan is mostly composed of PLC modems with lower power usage.

7 Discussion

In this section, we discuss the limitations of PowerMan and experiences from constructing and operating the PLC testbed and PowerMan prototype.

Interference in DCPS: The major contributor to the loss of performance (§3.2) is high-frequency noise from sources of interference in DCPS, including lighting, cooling, mechanical system, etc. As pointed out in [60], PLC is an extremely harsh environment for the high-bandwidth, high-frequency communication signals, as critical channel parameters (e.g. noise, impedance, and attenuation) are highly unpredictable [69]. Given the severity of the interference, in the design of PowerMan, we aim to limit the PLC signal within each PDU (i.e. within a PLC LAN (PLAN)) which reduces the signals exposure to interference, as the signal only travels short distances within the PDU. The LPF in each layer also removes high frequency noises from non-PLC sources.

Low Throughput Alternatives to PLC: Our experiments and simulations (6) exhibit low throughput for various traffic patterns, and as we have discussed, the reasons include noise, signal attenuation, MAC layer overhead, etc. Due to the low throughput of PowerMan, it is natural to consider using low cost, low bandwidth WiFi or Ethernet devices as alternatives. Compared to low cost alternatives, we believe PLC is advantageous in the three goals we outlined for an out-of-band MN (§1): survivability, deployability, and scalability.

- *Survivability:* PLC can leverage the robustness in existing power systems. Power system is the last-to-fail system in datacenters, and is independent of the data plane network. Embedded in DCPS, PowerMan can survive data plane failures, or even power system failures in Tier-II to IV datacenters, and is ready for immediate diagnosis and recovery. Other low cost alternatives do not share this quality. For example, a separate WiFi network requires additional monitoring and management systems to achieve the same level of robustness of power system.

[5] We use retail price here. The per-unit price are dependent on many factors: quantity, availability, distance, etc. With large quantity, the price tend to decrease.

[6] We use a rack-mounted server as the packet-forwarding hardware in the prototype, thus the high power usage. This can be reduced with a typical packet-forwarding device.

- *Deployability:* PowerMan reuses wiring in existing DCPS, thus there is no need to change ceiling height, and rack dimensions. This compatibility with existing datacenter designs greatly reduces the deployment cost. In contrast, WiFi-based solutions require changes in the rack dimensions to accommodate antennae of servers and access points. Ethernet-based solutions require additional rack space and cabling.
- *Scalability:* PowerMan reaches every device in the datacenter, as it reuses wiring in DCPS. Ethernet-based solutions with the same topology as PowerMan can reach the same port count, but at the cost of much more cabling. Like PLC, WiFi also suffers from the interference, which is more difficult to manage than that in a wired network. PowerMan is able to use signal filters on the border of two PLANs to eliminate interference between them. Such is not so easy in a wireless network, as there is no clear border between two broadcast domains. For WiFi-based solutions, handling interference requires careful planning of antennae direction, AP radio power, location, and channel selection. At the scale of a modern datacenter, the management of WiFi-based solution is challenging.

DC datacenters: Many modern datacenters are using DC power [31, 55, 68]. Our design can also work on such DCPS, because PowerMan is a design that utilizes power lines, which is the same in both DC and AC power systems. The carrier frequencies in PLC devices (assuming compliance with HomePlug standards) come from OFDM circuitries, and are not the 50-60Hz AC power.

Security Concerns: Datacenter MN is a high value target, and a MN using PLC may be vulnerable to on-premise attacks. PowerMan can adopt security mechanisms on MAC, network, transport, application layers. For example, in the PLC MAC layer, HomePlug 1.0 [56] supports 56-bit DES encryption, and later versions (HomePlug AV/AV2 [24, 81]) support 128-bit AES.

Cooling: Even with intensive experiments on PowerMan prototype, we have not yet witnessed any overheating issues for PLC modems. This is because: 1) the PLC modems have low power profile, and 2) the PLC modems are placed outside of the servers. Bump-in-the-Wire PSU design may benefit from the same reasons; but it is still important to investigate the heat dissipation of the Full-Integration design inside a rack-mounted server or switch as future work.

8 Related works

We summarize the related work in three broad categories: datacenter management, alternative datacenter networking architectures, and PLC networking.

Datacenter Management: There is vast literature on the management and control planes of datacenter networks [20, 42, 49, 52, 70, 85]. They often assume that the management traffic can be delivered, and PowerMan complements these works with an out-of-band MN that offers necessary latency and bandwidth, while being survivable, scalable, and deployable.

Datacenter Networking Architectures: Datacenter networks in production usually use the Clos network [19, 43, 54, 65, 67] to achieve high bisection bandwidth. Using flexible networking technology such as optical switching [32, 34, 38, 57, 59, 64, 80], FSO [41, 48], and 60GHz wireless radios [36, 47], dynamic network topologies are proposed to mitigate traffic hotspots and changing demands. We differ from them in our technology choice. In terms of datacenter MN, Angora [85] proposed using 60GHz wireless radio to construct a datacenter "facility network", which is a MN but with much stricter latency requirements. In contrast, PowerMan is the first attempt to employ PLC in the datacenter MN setting, and as our cost comparisons (§6.3) suggest, PowerMan has lower initial cost and operating power consumption than the other technologies at the same scale.

PLC Networking: In PLC PHY [60, 69] and MAC [51, 74, 75, 76, 77, 78] layer, many efforts have been made to improve the bandwidth, reliability, and latency [83]. In comparison, PowerMan focuses on the application of PLC in MN, exploring networking (§3.1) and scalability (§4.3) for datacenter management. PowerMan can benefit from all PHY and MAC layer optimizations (e.g. parameter setting, dynamic bandwidth allocation scheme), as they improve the PLANs in PowerMan.

9 Conclusion

This paper has introduced PLC as an out-of-band management channel for datacenters. We build a small-scale PLC testbed, and demonstrate the potential of PLC with deployment of actual management applications. In the process, we identified the wiring and scalability issues which prevent deployment of PLC in datacenters. To tackle these problems, we design PowerMan, a datacenter MN using PLC that can be implemented using commercially available PLC devices. We build a PowerMan prototype on a small testbed of 12 servers. Using experiments and large-scale simulations, we evaluate its performance, reliability, and cost-effectiveness.

For future work, we plan to 1) investigate custom PLC devices with optimized PHY/MAC layers to improve latency, throughput, scalability, and reliability; 2) integrate PSU with single-board computer, so as to provide isolation from local OS-related failures.

Acknowledgements: This work is supported in part by Hong Kong RGC ECS-26200014, GRF-16203715, GRF-613113, CRF-C703615G, & China 973 Program No.2014CB340303. We thank the anonymous NSDI reviewers and our shepherd Shyam Gollakota for their constructive feedback and suggestions.

References

[1] Ansible. https://www.ansible.com/. (Accessed on 01/08/2017).

[2] ansible/ansible-examples. https://github.com/ansible/ansible-examples/. (Accessed on 01/19/2017).

[3] Chef. https://www.chef.io/chef/. (Accessed on 01/08/2017).

[4] chef-web-docs/resource_examples.rst at master chef/chef-web-docs. https://github.com/chef/chef-web-docs/blob/master/chef_master/source/resource_examples.rst. (Accessed on 01/19/2017).

[5] Configuration openstack havana with ryu. https://github.com/osrg/ryu/wiki/configuration_openstack_havana_with_ryu. (Accessed on 01/08/2017).

[6] datacenter/empirical-traffic-gen: Simple client-server application for generating user-defined traffic patterns. https://github.com/datacenter/empirical-traffic-gen. (Accessed on 01/08/2017).

[7] Introduction to ad-hoc commands ansible documentation. http://docs.ansible.com/ansible/intro_adhoc.html. (Accessed on 01/20/2017).

[8] Netgear pl1200. https://www.netgear.com/home/products/networking/powerline/PL1200.aspx?cid=wmt_netgear_organic. (Accessed on 01/24/2017).

[9] Nginx. https://www.nginx.com/. (Accessed on 01/19/2017).

[10] nginx cookbook - chef supermarket. https://supermarket.chef.io/cookbooks/nginx. (Accessed on 01/19/2017).

[11] Openstack docs: Network design. http://docs.openstack.org/ops-guide/arch-network-design.html. (Accessed on 01/07/2017).

[12] osrg/ryu: Ryu component-based software defined networking framework. https://github.com/osrg/ryu. (Accessed on 01/08/2017).

[13] Product faq powerline adapters. http://kb.netgear.com/20233/Product-FAQ-Powerline-Adapters?cid=wmt_netgear_organic. (Accessed on 01/15/2017).

[14] Ryubook.pdf. https://osrg.github.io/ryu-book/en/Ryubook.pdf. (Accessed on 01/19/2017).

[15] Tia-942 telecommunications infrastructure set. https://global.ihs.com/tia_telecom_infrastructure.cfm?RID=Z56&MID=5280. (Accessed on 09/25/2017).

[16] Tp-link av1200 gigabit passthrough powerline starter kit. http://www.tp-link.com/ph/products/details/cat-18_TL-PA8010P-KIT.html. (Accessed on 01/24/2017).

[17] Trendnet powerline 1200 av2 adapter kit. https://www.trendnet.com/products/powerline-1200/TPL-420E2K. (Accessed on 01/24/2017).

[18] Data center site infrastructure tier standard: Topology. *Uptime Institute* (2012).

[19] AL-FARES, M., LOUKISSAS, A., AND VAHDAT, A. A scalable, commodity data center network architecture. In *ACM SIGCOMM* (2008).

[20] AL-FARES, M., RADHAKRISHNAN, S., RAGHAVAN, B., HUANG, N., AND VAHDAT, A. Hedera: Dynamic flow scheduling for data center networks. In *USENIX NSDI* (2010).

[21] ALIZADEH, M., GREENBERG, A., MALTZ, D. A., PADHYE, J., PATEL, P., PRABHAKAR, B., SENGUPTA, S., AND SRIDHARAN, M. Data center tcp (dctcp). *ACM SIGCOMM Computer Communication Review 41*, 4 (2011), 63–74.

[22] ALIZADEH, M., YANG, S., SHARIF, M., KATTI, S., MCKEOWN, N., PRABHAKAR, B., AND SHENKER, S. pfabric: Minimal near-optimal datacenter transport. In *ACM SIGCOMM* (2013).

[23] ALLIANCE, H. Homeplug_av2_whitepaper_20130909.pdf. https://www.codico.com/fxdata/codico/prod/media/Datenblaetter/AKT/HomePlug_AV2_whitepaper_20130909.pdf. (Accessed on 02/10/2018).

[24] ALLIANCE, H. Homeplug av specification. *Version 1*, 2006.12 (2007), 16.

[25] BAI, W., CHEN, L., CHEN, K., HAN, D., TIAN, C., AND SUN, W. Information-agnostic flow scheduling for commodity data centers. In *USENIX NSDI* (2015).

[26] BANERJEE, S., BHATTACHARJEE, B., AND KOMMAREDDY, C. *Scalable application layer multicast*, vol. 32. ACM, 2002.

[27] BARLOW, R. E., AND PROSCHAN, F. *Mathematical theory of reliability*. SIAM, 1996.

[28] BARROSO, L., DEAN, J., AND HOEZLE, U. Web search for a planet: the architecture of the google cluster. *IEEE Micro 23*, 2 (2003), 22–28.

[29] BARROSO, L. A., CLIDARAS, J., AND HÖLZLE, U. The datacenter as a computer: An introduction to the design of warehouse-scale machines. *Synthesis lectures on computer architecture 8*, 3 (2013), 1–154.

[30] BENSON, T., ANAND, A., AKELLA, A., AND ZHANG, M. Microte: Fine grained traffic engineering for data centers. In *CoNEXT* (2010).

[31] BORS, D. Data center power system design debate: Ac or dc? http://www.ecmweb.com/power-quality-archive/data-center-power-system-design-debate-ac-or-dc. (Accessed on 02/10/2018).

[32] CHEN, K., SINGLA, A., SINGH, A., RAMACHANDRAN, K., XU, L., ZHANG, Y., WEN, X., AND CHEN, Y. Osa: An optical switching architecture for data center networks with unprecedented flexibility. In *USENIX NSDI* (2012).

[33] CHEN, K.-C. Medium access control of wireless lans for mobile computing. *IEEE Network 8*, 5 (1994), 50–63.

[34] CHEN, L., CHEN, K., ZHU, Z., YU, M., PORTER, G., QIAO, C., AND ZHONG, S. Enabling wide-spread communications on optical fabric with megaswitch. In *NSDI* (2017).

[35] CISCO. Data center power and cooling. http://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/unified-computing/white_paper_c11-680202.pdf. (Accessed on 01/24/2017).

[36] CUI, Y., XIAO, S., WANG, X., YANG, Z., ZHU, C., LI, X., YANG, L., AND GE, N. Diamond: Nesting the data center network with wireless rings in 3d space. In *USENIX NSDI* (2016).

[37] DEAN, J. Software engineering advice from building large-scale distributed systems.

[38] FARRINGTON, N., PORTER, G., RADHAKRISHNAN, S., BAZZAZ, H. H., SUBRAMANYA, V., FAINMAN, Y., PAPEN, G., AND VAHDAT, A. Helios: A hybrid electrical/optical switch architecture for modular data centers. In *ACM SIGCOMM* (2010).

[39] FERREIRA, H. C., GROVE, H., HOOIJEN, O., AND VINCK, A. H. Power line communications: an overview. In *AFRICON, 1996., IEEE AFRICON 4th* (1996), vol. 2, IEEE, pp. 558–563.

[40] GALLI, S., SCAGLIONE, A., AND WANG, Z. For the grid and through the grid: The role of power line communications in the smart grid. *Proceedings of the IEEE 99*, 6 (2011), 998–1027.

[41] GHOBADI, M., MAHAJAN, R., PHANISHAYEE, A., DEVANUR, N., KULKARNI, J., RANADE, G., BLANCHE, P.-A., RASTEGARFAR, H., GLICK, M., AND KILPER, D. Projector: Agile reconfigurable data center interconnect. In *ACM SIGCOMM* (2016).

[42] GREENBERG, A., HAMILTON, J., MALTZ, D. A., AND PATEL, P. The cost of a cloud: research problems in data center networks. *ACM SIGCOMM computer communication review 39*, 1 (2008), 68–73.

[43] GREENBERG, A., HAMILTON, J. R., JAIN, N., KANDULA, S., KIM, C., LAHIRI, P., MALTZ, D. A., PATEL, P., AND SENGUPTA, S. VL2: A scalable and flexible data center network. In *ACM SIGCOMM* (2009).

[44] GRID, T. G. Power equipment and data center design. https://www.thegreengrid.org/en/resources/library-and-tools/382-Power-Equipment-and-Data-Center-Design. (Accessed on 01/24/2017).

[45] GUNGOR, V. C., AND LAMBERT, F. C. A survey on communication networks for electric system automation. *Computer Networks 50*, 7 (2006), 877–897.

[46] GUO, C., LU, G., LI, D., WU, H., ZHANG, X., SHI, Y., TIAN, C., ZHANG, Y., AND LU, S. BCube: A high performance, server-centric network architecture for modular data centers. In *SIGCOMM* (2009).

[47] HALPERIN, D., KANDULA, S., PADHYE, J., BAHL, P., AND WETHERALL, D. Augmenting data center networks with multi-gigabit wireless links. In *ACM SIGCOMM Computer Communication Review* (2011), vol. 41, ACM, pp. 38–49.

[48] HAMEDAZIMI, N., QAZI, Z., GUPTA, H., SEKAR, V., DAS, S. R., LONGTIN, J. P., SHAH, H., AND TANWERY, A. Firefly: A reconfigurable wireless data center fabric using free-space optics. In *ACM SIGCOMM* (2014).

[49] HASSAS YEGANEH, S., AND GANJALI, Y. Kandoo: a framework for efficient and scalable offloading of control applications. In *ACM HotNets* (2012).

[50] HENDERSON, T. R., LACAGE, M., RILEY, G. F., DOWELL, C., AND KOPENA, J. Network simulations with the ns-3 simulator. *SIGCOMM demonstration 15* (2008), 17.

[51] HENRI, S., VLACHOU, C., HERZEN, J., AND THIRAN, P. Empower hybrid networks: Exploiting multiple paths over wireless and electrical mediums. In *ACM CoNEXT* (2016).

[52] ISARD, M. Autopilot: automatic data center management. *ACM SIGOPS Operating Systems Review 41*, 2 (2007), 60–67.

[53] KALBFLEISCH, J. D., AND PRENTICE, R. L. *The statistical analysis of failure time data*, vol. 360. John Wiley & Sons, 2011.

[54] KASSING, S., VALADARSKY, A., SHAHAF, G., SCHAPIRA, M., AND SINGLA, A. Beyond fat-trees without antennae, mirrors, and disco-balls. In *ACM SIGCOMM* (2017).

[55] KASSNER, M. P. Dc distribution is not just for the giants. http://www.datacenterdynamics.com/content-tracks/design-build/dc-distribution-is-not-just-for-the-giants/95037.fullarticle. (Accessed on 02/10/2018).

[56] LEE, M., NEWMAN, R. E., LATCHMAN, H. A., KATAR, S., AND YONGE, L. Homeplug 1.0 powerline communication lans: protocol description and performance results. *International Journal of Communication Systems 16*, 5 (2003), 447–473.

[57] LIU, Y. J., GAO, P. X., WONG, B., AND KESHAV, S. Quartz: a new design element for low-latency dcns. In *ACM SIGCOMM* (2014).

[58] MCKEOWN, N., ANDERSON, T., BALAKRISHNAN, H., PARULKAR, G., PETERSON, L., REXFORD, J., SHENKER, S., AND TURNER, J. Openflow: Enabling innovation in campus networks. *ACM Computer Communication Review* (2008).

[59] MELLETTE, W. M., MCGUINNESS, R., ROY, A., FORENCICH, A., PAPEN, G., SNOEREN, A. C., AND PORTER, G. Rotornet: A scalable, low-complexity, optical datacenter network. In *ACM SIGCOMM* (2017).

[60] MENG, H., CHEN, S., GUAN, Y., LAW, C., SO, P., GUNAWAN, E., AND LIE, T. Modeling of transfer characteristics for the broadband power line communication channel. *IEEE Transactions on Power delivery 19*, 3 (2004), 1057–1064.

[61] NEE, R. V., AND PRASAD, R. *OFDM for wireless multimedia communications*. Artech House, Inc., 2000.

[62] PERRY, J., OUSTERHOUT, A., BALAKRISHNAN, H., SHAH, D., AND FUGAL, H. Fastpass: A centralized zero-queue datacenter network. In *ACM SIGCOMM* (2014).

[63] PFAFF, B., PETTIT, J., KOPONEN, T., JACKSON, E. J., ZHOU, A., RAJAHALME, J., GROSS, J., WANG, A., STRINGER, J., SHELAR, P., ET AL. The design and implementation of open vswitch. In *USENIX NSDI* (2015).

[64] PORTER, G., STRONG, R., FARRINGTON, N., FORENCICH, A., SUN, P.-C., ROSING, T., FAINMAN, Y., PAPEN, G., AND VAHDAT, A. Integrating microsecond circuit switching into the data center. In *ACM SIGCOMM* (2013).

[65] ROY, A., ZENG, H., BAGGA, J., PORTER, G., AND SNOEREN, A. C. Inside the social network's (datacenter) network. In *ACM SIGCOMM* (2015).

[66] SCHWARTZ, M. Carrier-wave telephony over power lines: Early history [history of communications]. *IEEE Communications Magazine 47*, 1 (2009), 14–18.

[67] SINGH, A., ONG, J., AGARWAL, A., ANDERSON, G., ARMISTEAD, A., BANNON, R., BOVING, S., DESAI, G., FELDERMAN, B., GERMANO, P., ET AL. Jupiter rising: A decade of clos topologies and centralized control in google's datacenter network. In *ACM SIGCOMM* (2015).

[68] STARK, J. 380v dc power: Shaping the future of data center energy efficiency. http://www.datacenterknowledge.com/archives/2015/06/25/380v-dc-power-shaping-future-data-center-energy-efficiency. (Accessed on 02/10/2018).

[69] TANG, L., SO, P., GUNAWAN, E., CHEN, S., LIE, T., AND GUAN, Y. Characterization of in-house power distribution lines for high-speed data transmission. In *Proc. 5th Int. Power Engineering Conf.(IPEC 2001)* (2001), pp. 7–12.

[70] TOOTOONCHIAN, A., AND GANJALI, Y. Hyperflow: A distributed control plane for openflow. In *Proceedings of the 2010 internet network management conference on Research on enterprise networking* (2010), pp. 3–3.

[71] TURNER IV, W. P., PE, J., SEADER, P., AND BRILL, K. Tier classifications define site infrastructure performance. *Uptime Institute* (2006).

[72] VAMANAN, B., HASAN, J., AND VIJAYKUMAR, T. Deadline-aware datacenter tcp (d2tcp). *ACM SIGCOMM Computer Communication Review* (2012).

[73] VLACHOU, C. plc-click-elements/plcstats.h. https://github.com/christinavl/plc-click-elements/blob/master/PLCStats.h. (Accessed on 02/10/2018).

[74] VLACHOU, C., BANCHS, A., HERZEN, J., AND THIRAN, P. Analyzing and boosting the performance of power-line communication networks. In *ACM CoNEXT* (2014).

[75] VLACHOU, C., BANCHS, A., HERZEN, J., AND THIRAN, P. On the mac for power-line communications: Modeling assumptions and performance tradeoffs. In *IEEE ICNP* (2014).

[76] VLACHOU, C., BANCHS, A., HERZEN, J., AND THIRAN, P. Performance analysis of mac for power-line communications. *ACM SIGMETRICS Performance Evaluation Review 42*, 1 (2014), 585–586.

[77] VLACHOU, C., BANCHS, A., SALVADOR, P., HERZEN, J., AND THIRAN, P. Analysis and enhancement of csma/ca with deferral in power-line communications. *IEEE Journal on Selected Areas in Communications* (2016).

[78] VLACHOU, C., HERZEN, J., AND THIRAN, P. Fairness of mac protocols: Ieee 1901 vs. 802.11. In *Power Line Communications and Its Applications (ISPLC), 2013 17th IEEE International Symposium on* (2013), IEEE, pp. 58–63.

[79] VLACHOU, C., HERZEN, J., AND THIRAN, P. Simulator and experimental framework for the mac of power-line communications. EPFL-REPORT-205770.

[80] WANG, G., ANDERSEN, D., KAMINSKY, M., PAPAGIANNAKI, K., NG, T., KOZUCH, M., AND RYAN, M. c-Through: Part-time optics in data centers. In *ACM SIGCOMM* (2010).

[81] YONGE, L., ABAD, J., AFKHAMIE, K., GUERRIERI, L., KATAR, S., LIOE, H., PAGANI, P., RIVA, R., SCHNEIDER, D. M., AND SCHWAGER, A. An overview of the homeplug av2 technology. *Journal of Electrical and Computer Engineering 2013* (2013).

[82] ZATS, D., DAS, T., MOHAN, P., BORTHAKUR, D., AND KATZ, R. Detail: reducing the flow completion time tail in datacenter networks. *ACM SIGCOMM Computer Communication Review 42*, 4 (2012), 139–150.

[83] ZHAO, Z., CHEN, I., ET AL. Moving homeplug to industrial applications with power-line communication network.

[84] ZHOU, X., ZHANG, Z., ZHU, Y., LI, Y., KUMAR, S., VAHDAT, A., ZHAO, B. Y., AND ZHENG, H. Mirror mirror on the ceiling: Flexible wireless links for data centers. *ACM SIGCOMM Computer Communication Review 42*, 4 (2012), 443–454.

[85] ZHU, Y., ZHOU, X., ZHANG, Z., ZHOU, L., VAHDAT, A., ZHAO, B. Y., AND ZHENG, H. Cutting the cord: a robust wireless facilities network for data centers. In *ACM MobiCom* (2014).

Appendix

Management Traffic Optimizations

As discussed in §3.2.2, two common traffic patterns of management application is 1-to-N (e.g. configuration tasks) and N-to-1 (e.g. monitoring tasks). The experiments in §3.2&6.1 shows that such patterns perform poorly on PowerMan. In the following, we propose application-layer traffic optimizations to reduce the completion times of these two patterns on PowerMan. We assume the controller is located at the root of the tree.

Accelerating 1-To-N Pattern

As shown in Figure 16, PowerMan has low per-server bandwidth at large scale. This is because the controller node needs to maintain connection to all servers, so the per-server bandwidth is constrained by the interface capacity of the controller. Low per-server bandwidth can prolong completion times of configuration tasks.

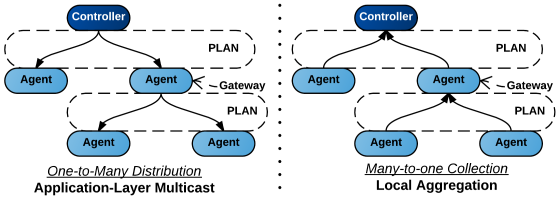


Figure 19: Accelerating Management Application Traffic Patterns

We propose to construct an application-layer multicast (ALM) overlay network [26] in PowerMan for management tasks with 1-to-N distribution pattern, where the gateway in each PDU act as a distribution agent in corresponding PLAN. As shown in Figure 19, the distribution from controller to all server is divided into multiple distributions within different PLANs.

We evaluate the performance of ALM. We use the production traces in §3.2.2. For baseline performance, we create 1-to-N traffic patterns using the traffic generator with different numbers of receivers. For ALM, we modify the traffic generator to include an implementation of ALM agent, and enable the agents in the gateways. We collect the FCTs and the results are plotted in Figure 20. We observe that ALM reduces the FCT for 1-to-N pattern, and the performance gap increases with the number of total receivers. For 10 receivers, the FCT is reduced by 15.38% on average. The main reason is that the total traffic volume is reduced with ALM, as copies of the flow are created by the agent in each gateway, which reduces the traffic volume at higher layers.

Accelerating N-to-1 Pattern

Information collection tasks include monitoring, diagnostics, and measurement, which exhibit N-to-1 pattern

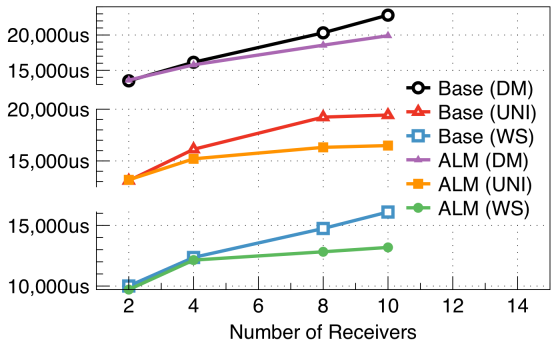


Figure 20: Accelerate 1-to-N Pattern with Application-Layer Multicast: Average FCT

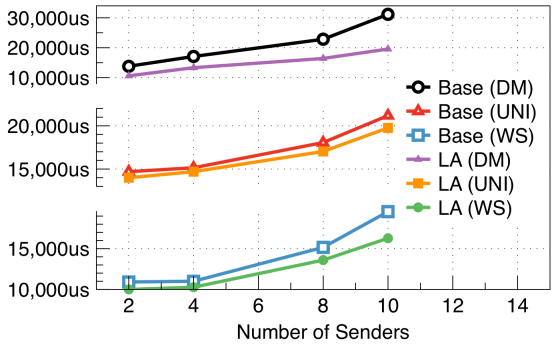


Figure 21: Accelerate N-to-1 Pattern with Local Aggregation: Average FCT

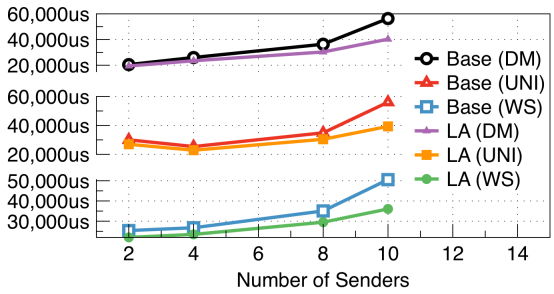


Figure 22: Accelerate N-to-1 Pattern with Local Aggregation: 99 percentile FCT

from the servers to the controller. As the dual of ALM, we propose local aggregation (LA) at each layer (PLAN) of the tree. An agent at each gateway collects the information from all servers/agents in its PLAN, and then sends the aggregated information to the agent in upper-layer gateway.

We then evaluate the performance of LA. For baseline, we create N-to-1 patterns using the same flow size distributions as above. For LA, we implemented a LA agent for the traffic generator, and enable them on all gateways. We collect the FCTs and plot the average in Figure 21 with respect to the number of senders. We can see that, although LA in general outperforms baseline, the performance gap is smaller than that of Figure 20. This is because the total traffic volume is not reduced with LA. However, LA on PowerMan effectively reduces the num-

ber of contending flows at the controller from N (total number of servers) to $k-1$ (number of nodes in a PLAN). This can be observed in Figure 22, which summarizes the tail latencies (99th percentile FCTs). Tail latencies capture the worst performing flows whose completion times are prolonged by events such as packet loss, reordering, frame collision. Using LA to reduce the number of contending flows at the receiver decreases the occurrences of tail latency events, which improves the completion times. Finally, a further optimization is to compress local information before sending. Compression of locally collected information can reduce total traffic volume, and it would be beneficial if the computation overhead on the gateway is acceptable.