IEEE CLOUD COMPUTING

INTELLIGENCE IN THE CLOUD

MapReduce-Based Ensemble Learning, p. 38

Intelligent Resource Management in Blockchain-Based Cloud Datacenters, p. 50

NOVEMBER/DECEMBER 2017
VOLUME 4, NUMBER 6
www.computer.org/cloudcomputing

IEEE
IEEE computer society
IEEE ComSoc
IEEE Communications Society

The 10 Laws of Fogonomics

Joe Weinman

Although fog computing goes by various names, it can be simply viewed as inserting one or more highly distributed compute and storage layers, possibly available on a pay-per-use basis, that are networked to centralized cloud data centers and, via an edge layer, to highly dispersed, often mobile, endpoints spanning user devices—such as smartphones and tablets—and things—such as smart electric meters and connected cars—as shown in Figure 1. This can be viewed as replacing what had been merely a communications channel—such as the Internet or direct connections—between the cloud and devices/things with a multilayered compute, storage, and network fabric.[1]

There are various tradeoffs between architecture choices such as cloud-endpoint or cloud-fog/edge-endpoint. For example, processing at the edge is closer to where data is generated or resides in user devices and things—such as video or image capture—but is farther from data residing in the cloud—such as web search indices/repositories. The cloud, which has hyperscale data centers, offers enormous capacity, but at the expense of latency and backhaul costs for interactive tasks supporting devices and things. These characteristics can be expressed both quantitatively and qualitatively. As a result, the economics of the cloud—*Cloudonomics*—can be differentiated from the economics of the fog—*Fogonomics*.

EDITOR
JOE WEINMAN
joeweinman@gmail.com

The 10 Laws of Cloudonomics

In 2008, I wrote “The 10 Laws of Cloudonomics,” quantifying the essential characteristics of cloud computing such as pay-per-use and on-demand.[2] This was expanded into the book *Cloudonomics: The Business Value of Cloud Computing*.[3] Briefly, the laws are:

1. **Utility services cost less even though they cost more**—even if the unit costs of public cloud resources are higher than private, pay-per-use pricing in the presence of variable demand can make the total cost lower. Depending on differences in pricing, workload demand variability, and performance differentials, hybrid clouds can often minimize total costs.

2. **On-demand trumps forecasting**—no matter how good workload demand forecasting is, near-real-time provisioning of resources will be better at exactly matching capacity with demand. This can only be economically done in a public cloud with dynamically allocated, shared resources.

2325-6095/17/$33.00 © 2017 IEEE

3. **The peak of the sum is never greater than the sum of the peaks**—so the aggregate capacity needed in a dynamically shared resource pool is typically less, and in the worst case, equal, to the capacity needed when divided into siloed private clouds.
4. **Aggregate demand is smoother than individual**—in other words, the coefficient of variation (the ratio of the standard deviation to the mean) of a sum of multiple independent, identically distributed random variables (representing individual workload demand levels) with nonzero means and variances is less than that of any individual one.
5. **Average unit costs are reduced by distributing fixed costs over more units of output**—in other words, economies of scale apply to public cloud providers operating datacenters with hundreds of thousands of servers.
6. **Superiority in numbers is the most important factor in the result of a combat – *Carl von Clausewitz***—public clouds have the size to be minimally affected by cyberattacks such as large botnets generating massive distributed denial-of-service attack bandwidth.
7. **Space-time is a continuum – *Albert Einstein/Hermann Minkowski***—embarrassingly parallel applications or tasks (say, the Map phase in MapReduce jobs) can trade off the number of processors for the compute time. In the presence of pay-per-use pricing, this means that acceleration is free.
8. **Dispersion is the inverse square of latency**—it takes four times as many nodes to reduce latency by half on a surface such as a plane, so eventually latency reduction becomes prohibitively expensive.
9. **Don't put all your eggs in one basket**—while a single enterprise data center is at risk of a smoking hole disaster, and even a nearby sister site may be taken out by the same disaster, say, a flood, tornado, or hurricane, a cloud offering many availability zones and regions each with good reliability can create a system architecture with excellent reliability.
10. **An object at rest tends to stay at rest – *Isaac Newton***—which captures the advantages of new, on-net sites in areas with cheap power and acreage over existing data centers which would be too costly to move or for which Power Usage Effectiveness improvements might have too long a payback period.

THE 10 LAWS OF CLOUDONOMICS

1. Utility Services Cost Less Even Though They Cost More
2. On-Demand Trumps Forecasting
3. The Peak of The Sum Is Never Greater Than the Sum of the Peaks
4. Aggregate Demand Is Smoother Than Individual
5. Average Unit Costs Are Reduced by Distributing Fixed Costs Over More Units of Output
6. Superiority in Numbers Is the Most Important Factor in The Result of a Combat
7. Space-Time Is a Continuum
8. Dispersion Is the Inverse Square of Latency
9. Don't Put All Your Eggs in One Basket
10. An Object at Rest Tends to Stay at Rest

The 10 Laws of Fogonomics

Some of the same insights apply to fog/edge computing, but some are less important or even irrelevant. For example, pay-per-use pricing *might* be used for edge resources, but is less central to the business and operating model of the edge than it is to public clouds. This is partly because public clouds can offer pay-per-use because they dynamically allocate a large shared resource pool across multiple different customers/workloads over time. With that in mind, we can highlight the important economic characteristics of fog computing, which in some cases are the same or similar to the economics of the cloud, but often different.

Fogonomics Law #1: Time Is of The Essence

Consider a world with only one data center, whether cloud, colocation, or enterprise. Let's assume that

THE 10 LAWS OF FOGONOMICS

1. Time Is of the Essence
2. Time Is Money
3. No Man Is an Island Entire of Itself
4. Don't Put All Your Eggs in One Basket
5. Divide and Conquer
6. United We Stand
7. Many Hands Make Light Work
8. A Bad Penny Always Turns Up
9. Space-Time Is a Continuum
10. Penny Wise, Pound Foolish

it is in the New York City area. This would provide excellent response time for New York Stock Exchange applications or for interactive applications such as online gaming for Manhattan residents. However, to service say, the Shanghai Stock Exchange or gamers living in Sydney, the response time would be horrible, given that the network round-trip time alone is on the order of 160 milliseconds. Consequently, latency-sensitive applications intended to serve global audiences need to either run within the device or at or near the edge of the fog. A number of studies have shown the economic benefits of reduced latency. For example, an additional 1/2 second delay in returning search results led to 20% fewer click-throughs, and thus a 20% drop in advertising revenues.[4–5]

Because the area within a radius r of a service node is πr^2, and either worst case or average latency l is proportional to that radius r, for n nodes the area A covered follows $A \propto n\pi r^2$ and thus $A \propto n\pi l^2$ and therefore, $l \propto \frac{1}{\sqrt{n}}$. There are some fine points due to packing density η (eta), irregularly shaped masses (such as continents), propagation and framing delays for various network media and protocols, the curvature of the earth, and non-great-circle network routes, but the point is that dispersing edge resources greatly reduces latency to/from endpoints and thus total response time (see Figure 2).

Fogonomics Law #2: Time Is Money – Antiphon

Not only do distributed resources reduce latency to and from devices and things, they are more efficient in terms of backhaul network transport capacity. The costs of a one-way journey or round-trip to the cloud can be expressed not just in terms of time and latency, but money: the capital investments in network capacity and/or charges for network transport. For example, a 5 MB picture taken with a smartphone that is kept on the smartphone uses no network resources, whereas one uploaded to the cloud does.

Network infrastructure costs can be difficult to quantify.[6] For example, the marginal cost to transport a packet may be zero if the network is uncongested, yet network infrastructure requires massive capital expenditures to deploy. Moreover, the cost and effort to dig trenches and to lay optical fiber are always high, the cost of a fiber is proportional to distance essentially regardless of how many waves are carried on it, and optoelectronics costs increase with data transmission rate, but do so sublinearly. Nevertheless, if we quantify network costs in terms of data traffic transported per mile, we can easily relate service node dispersion to cost reduction of data transport over network infrastructure.

As stated in Law #1, for any given area A, latency l and the number of service nodes n essentially follow $l \propto \frac{1}{\sqrt{n}}$. Moreover, the distance d that data needs to be transported is (roughly) proportional to the latency l. There are a number of reasons that it isn't exactly proportional, such as the fact that physical wireline network routes don't necessarily follow the shortest path, but are dependent on things such as rights-of-way on railroads and thus are artifacts of those routes.

Consequently, an increase in service node dispersion from n_c cloud nodes to n_f fog/edge nodes leads to a reduction in transport capacity-miles needed to carry a volume of data V in a given time period T of $\frac{V}{T}\frac{1}{\sqrt{n_c}} - \frac{V}{T}\frac{1}{\sqrt{n_f}}$.

Fogonomics Law #3: No Man Is an Island Entire of Itself – *John Donne*

Of course, adding more nodes can mean adding more interconnections, although the exact increase depends on network topology and architecture, as shown in Figure 3.

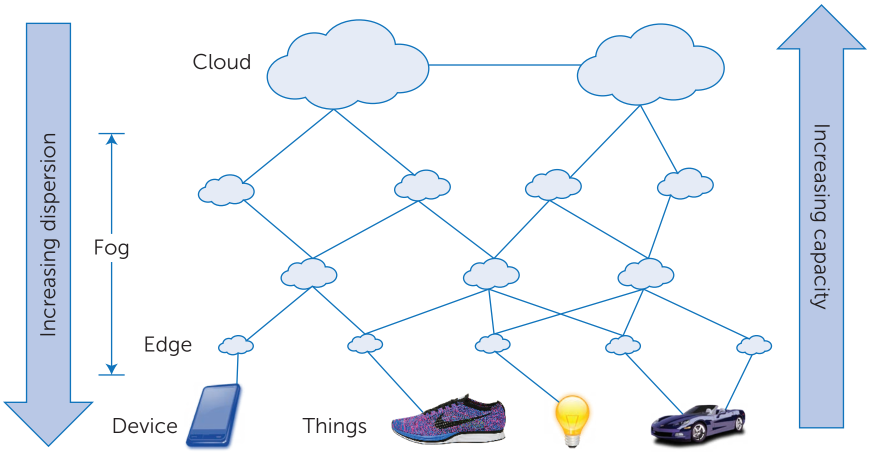

FIGURE 1. Cloud, Fog, Edge, Devices, and Things.

For example, if d devices and/or things were all individually connected to a single cloud data center via dedicated point-to-point connections, there would be d connections. If each device and/or thing is connected to exactly one of n_f fog/edge nodes, there are still d connections there, but then there are also n_f connections from these nodes to the cloud data center, and possibly $n_f(n_f - 1)/2$ connections between fog/edge nodes if they are fully connected. If each device or thing is connected to two or more fog/edge nodes for reliability, and those nodes are connected to the cloud and to each other, then the total number of connections increases to $2d + n_f + n_f(n_f - 1)/2$, etc. The cost of interconnections varies depending on whether they are wireless or wireline, physical or virtual, their bandwidth requirements, etc.

However, the growth in the number of connected things and online users and their devices is real, leading to quantifiable increases in interconnections and "Interconnection Bandwidth." It has been characterized in a recent analysis by interconnection and colocation company Equinix, via their Global Interconnection Index.[7]

Fogonomics Law 4: Don't Put All Your Eggs in One Basket

As with the corresponding law of Cloudonomics, system reliability and availability can improve through replication, depending on the failure mode. For independent node failures due to localized physical phenomena such as failure of a critical component representing a single point of failure, an overall system is unlikely to experience total failure, but may experience limited or negligible performance degradation.

In any given time period, if the probability of failure of any given one of n_f fog/edge nodes is f, then the probability that the node is available is $(1 - f)$. Assuming that node failures are independent, the probability that the overall system is completely down is f^{n_f}. Therefore, the probability that there is still some functionality is $1 - f^{n_f}$. As the number of nodes n_f increases, f^{n_f} approaches 0, so $1 - f^{n_f}$ approaches unity and thus at least part of the system can be expected to function.

On the other hand, sometimes failures are not independent and localized, but systemic, for example due to system-wide software issues, design issues, or cascading faults. In this case, the quantity of nodes makes no difference. In the public cloud domain, consider the Netflix Christmas Eve outage, due to AWS Elastic Load Balancer control-plane issues.[8] In fact, having more nodes could increase the costs of repair and recovery: consider defects outside of the cloud/fog domain such as with the Intel Pentium chip floating point design flaw or the Takata airbag recall.

Fogonomics Law #5: Divide and Conquer – *Philip II of Macedon*

Division of *processing* resources has a number of benefits over consolidation

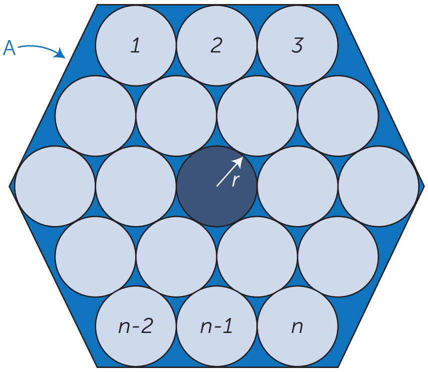

FIGURE 2. The relationship between service nodes n and radius r for a fixed area A.

of those resources, as discussed above. But what about *storage*? Ultimately it depends on the exact application architecture. For example, consider a retail chain that maintains point of sale data from any given store at that store or at one single nearby fog/edge node. Or, consider a video surveillance application that maintains data locally, unless a certain movement threshold is reached, at which point it uploads it to the cloud. Data is then partitioned, and except for quantization effects, such as minimum storage quantity at each node, the total data storage requirements remain

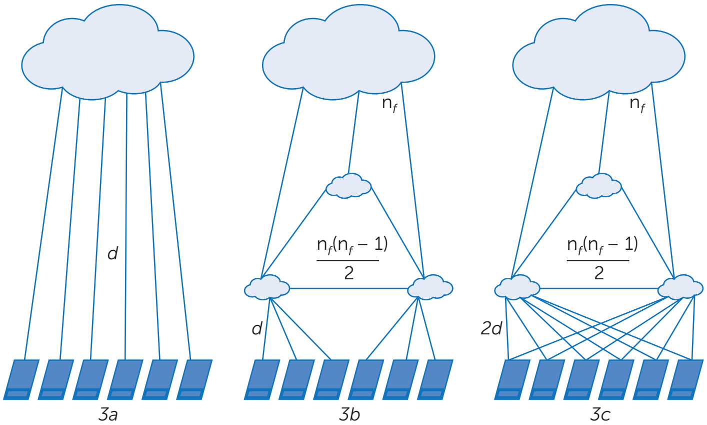

FIGURE 3. Growth in interconnection.

unchanged if the data is partitioned. In other words, rather than one instance of storage of size S, we have an equivalent need for $n_f \times (S/n_f)$, or, if there is a minimum data storage resource size m, a need for $n_f \times (\max(m, S/n_f))$.

At the other extreme, the total amount of storage might be proportional to the number of nodes. For example, a bill of materials might be maintained at each factory producing a given product. Or, instead of data storage we might consider storage requirements for the image of the application and/or operating system(s). In this case, replication and dispersion of resources into n_f nodes lead to a multiplication of storage requirements by n_f to $n_f \times S$.

Fogonomics Law #6: United We Stand – *John Dickinson*

In the last issue of *IEEE Cloud Computing* magazine, I looked in depth at capacity, cost, and utilization benefits to resource aggregation.[9] Because fog stands between the cloud and devices/things, it can create benefits—or issues—in terms of total capacity requirements. Consolidated cloud resources have benefits in terms of total capacity requirements over partitioned fog resources. On the other hand, resources consolidated at the fog layer have benefits in terms of total capacity requirements over partitioned devices or things.

As a quick recap, assume that there is a workload whose level of demand in some unit (say, t2.micro or m4.xlarge) varies as a Normal random variable with mean μ and variance σ^2. As an example, if we want to make sure that we have enough capacity at least 97.7% of the time, we need to set the capacity to be at least $\mu + 2\sigma$ when we are running that workload in a siloed environment. For different expected availability targets, we need to set the capacity to different levels, say, $\mu + k\sigma$. If we run n such workloads each in their own siloed environment, we need $n(\mu + k\sigma) = n\mu + kn\sigma$ units of capacity. If we aggregate multiple such workloads, and their demands are independent, there is a statistical smoothing effect, namely, a reduction in the coefficient of variation. The variance of the sum is the sum of the variances, namely $n\sigma^2$, so the standard deviation of the sum is $\sqrt{n}\sigma$. Thus, to have the same degree of sufficient capacity, we would only need $n\mu + k\sqrt{n}\sigma$ total capacity. Since for $n > 1$ the square root of n is smaller than n, this means that the aggregate capacity requirements for achieving the same level of service availability are lower. So, if resources are united, we stand to reduce total capacity requirements needed to achieve a given level of expectation of sufficient capacity.

Fogonomics Law #7: Many Hands Make Light Work – *John Heywood*

As the saying goes, many hands make light work. Unfortunately, light work is bad if you are measuring utilization. Here again, cloud trumps fog, but fog trumps devices/things. Specifically, the expected value of the total amount of work across n workloads each with mean μ is simply $n\mu$. Doing the same amount of work with fewer resources means that average utilization will be higher. Specifically, the utilization of partitioned workloads each running in its own siloed resources is $n\mu/(n\mu + kn\sigma)$, whereas the utilization of aggregate, dynamically allocated, shared capacity by those same workloads is higher, namely $n\mu/(n\mu + k\sqrt{n}\sigma)$.[9] As shown in Figure 4, for any given target k, utilization levels from shared resource pooling get better as more workloads are aggregated (moving to the right), or conversely, worse as fewer workloads are (moving to the left).

Fogonomics Law #8: A Bad Penny Always Turns Up

There is one final implication of the statistics of aggregation into or out of the fog having to do with cost. High or low utilization per se may not be viewed as critical, but it *does* have an economic implication. Specifically, a bad cost structure will turn up either in the price at which services are delivered relative to competitors, *or* in reduced profitability. While the total cost structure of any computing architecture, whether delivered as a service or not depends on many factors having nothing to do with

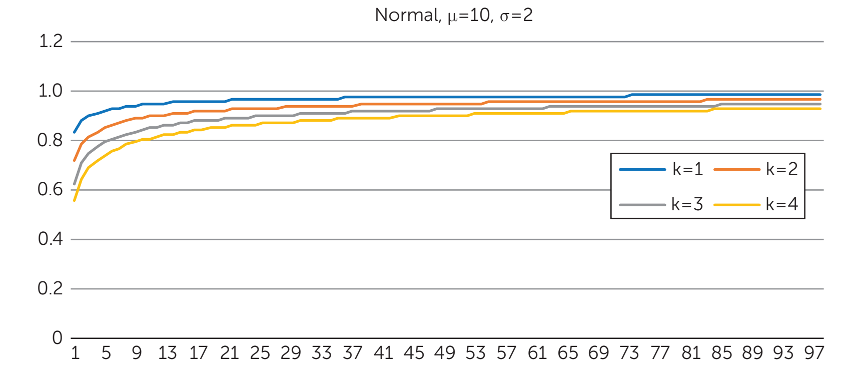

FIGURE 4. Utilization $U = f(n, \mu, \sigma, k)$ where $\mu = 10$ and $\sigma = 2$ for various k from $n =$ 1 to 100.

that architecture (e.g., the Chief Executive Officer's bonus, lawsuit settlements, and tax credits) it is clear that no matter what the total is, it's better—for either price, profitability, or both—to have a better cost structure, all other things being equal.

In the case of resource aggregation, poor utilization means that those entities that are paying for resources are not only paying for *used* resources, but will also have to pay, one way or another, for *unused* resources. The increase in cost structure for at least the physical resource component of the computing resources can then be quantified as a premium over the cost of the resources that are actually utilized. In the case of siloed resources, the total amount of resources needed is $n(\mu + k\sigma)$. In the case of aggregated resources (again, supporting independent, uncorrelated demand), the total needed is only $(n\mu + k\sqrt{n}\sigma)$.[9] Therefore, we need to include a premium of $n(\mu + k\sigma)/(n\mu + k\sqrt{n}\sigma)$ when determining a cost structure for sold resources to recapture the cost of unused capacity.

Fogonomics Law #9: Space-Time Is a Continuum – *Albert Einstein/ Hermann Minkowski*

As highlighted in Cloudonomics Law #7, space (i.e., number of processors) can be traded off against time (i.e., time to execute a compute job). Thus, an embarrassingly parallel workload running on one CPU might run 100 times faster on 100 CPUs, say in 1 hour rather than 100 hours. In a cloud environment with pay-per-use pricing, this means that acceleration is free, because in either case the user will be charged for 100 CPU hours.

The fog supports parallelism well, subject to internode communications costs, but one issue with the fog is that if resources are dedicated and without pay-per-use pricing, this type of benefit vanishes. Pay-per-use pricing is economically feasible for a service provider only because of the statistical benefits described in some of the above laws. Without resource pooling and dynamic resource allocation to time-varying workloads, pay-per-use is uneconomical. As a simple analogy, Avis rental car services couldn't enjoy sustained economic viability if they were to charge *you* a 30-dollar day rate for a car for *only* one day but reserve the car *only* for you for three years, and tuck it away in the back of the garage at other times.

For some applications, we can treat the fog as a highly parallel distributed computing medium, for example, weather sensors. The key considerations for whether dispersed fog or hyperscale cloud are relevant include the nature of the application and the degree of real-time interprocessor communication. Some applications, say, sensor data collection with data transfer only when alarm thresholds are passed, typically will have little interprocessor communication. On the other hand, a neuromorphic deep learning application may have a massive amount of such communication, and be better suited to, say, a nondispersed hypercube architecture.[10]

Fogonomics Law #10: Penny Wise, Pound Foolish – *Edward Topsell*

Notwithstanding the quantitative benefits of the above 9 laws, we can't ignore the strategic reality and benefits of emerging fog/edge architectures. The digitalization of organizations, consumers, processes, products, services and their increasing ubiquity inherently is driving edge functionality. For example, smart, digital, connected door locks are battery-operated due to the nature of their physical implementation. The trade-offs between size, signal strength, and battery life drive low-power networking solutions such as Z-Wave, in turn driving a need for a hub or gateway to connect to the Internet. In other words, consumer needs and revenue growth objectives of smart home vendors inarguably drive a fog/edge architecture.

In addition to these 10 laws, there are other application-dependent benefits of the fog. For example, when devices and things collect data, machine learning and deep learning algorithms can process that data to develop inferences and correlations.[11] As a rule, the more data there is, the greater the ability to surface inferences and the greater the confidence in the inference. However, at some point, there can be diminishing returns to the quality of the insights generated relative to the quantity of data collected and analyzed.

Autonomy, privacy, sovereignty, and security can also be benefits of the fog/edge. For example, a manufacturing-oriented fog must *autonomously* keep the factory up and running by sensing and controlling various factory elements such as robots and continuous processes, whether or not a central cloud is suffering an outage. Certain collected data may be kept *private* and/or *secure*, never being sent to a distant location. And, certain countries mandate that IT comply with data *sovereignty* laws and regulations, preventing, say, medical data from ever crossing the border.

Summary

Fog/edge computing represents an emerging approach that has a number of economic benefits, but also some weaknesses relative to traditional cloud computing. For example, latency and backhaul bandwidth requirements are substantially reduced for transactions regarding endpoints such as user devices and smart, connected things. On the other hand, there are some benefits to a cloud architecture, such as workload aggregation and the ability to run large workloads with highly interconnected micro-services or tasks. Both approaches are likely to coexist in a hybrid fashion: the hybrid, multilayer cloud-fog-edge architecture.

References

1. Y. Yang, “FA2ST: Fog As A Service Technology,” prepublication draft.
2. J. Weinman, “The 10 Laws of Cloudonomics,” *Gigaom.com*, 7 Sept. 2008; https://gigaom.com/2008/09/07/the-10-laws-of-cloudonomics/.
3. J. Weinman, *Cloudonomics: The Business Value of Cloud Computing*, Wiley, 2012.
4. D. Farber, “Google’s Marissa Mayer: Speed Wins,” *ZDnet.com*, 9 Nov. 2006; http://www.zdnet.com/article/googles-marissa-mayer-speed-wins/.
5. J. Weinman, “The Cloud and the Economics of User Experience,” *IEEE Cloud Computing*, vol. 2, no. 6, 2015, pp. 74–78.
6. H. Wagter, “Fiber-to-the-X: The Economics of Last-Mile Fiber,” *Ars Technica*, 30 Mar. 2010; https://arstechnica.com/tech-policy/2010/03/fiber-its-not-all-created-equal/.
7. Equinix, “Global Interconnection Index, 2017”; https://www.equinix.com/interconnection-enables-the-digital-economy/.
8. “Summary of the December 24, 2012 Amazon ELB Service Event in the US-East Region”; https://aws.amazon.com/message/680587/.
9. J. Weinman, “The Economics of Computing Workload Aggregation: Capacity, Utilization, and Cost Implications,” *IEEE Cloud Computing Mag.*, vol. 4, no. 5, pp. 6–11.
10. J. Błazewicz and M. Drozdowski, “Scheduling Divisible Jobs on Hypercubes,” *Parallel Computing*, vol. 21, no. 12, Dec. 1995, pp. 1945–1956.
11. A. Morshed et al., “Deep OSMOSIS: Holistic Distributed Deep Learning in Osmotic Computing”, *IEEE Cloud Computing*, Nov./Dec. 2017.

JOE WEINMAN *is a frequent global keynoter and author of Cloudonomics and Digital Disciplines. He has held executive leadership positions at AT&T, HP, and Telx. He also serves on the advisory boards of several technology companies. Weinman has a BS in computer science from Cornell University and an MS in computer science from the University of Wisconsin-Madison. He has completed executive education at the International Institute for management Development in Lausanne. Weinman has been awarded 22 patents. Contact him at joeweinman@gmail.com*