Latent Effects of Cloud Computing on IT Capacity Management Structures

Joe Bauer*, Al Bellamy
Eastern Michigan University, Ypsilanti, MI, USA.

* Corresponding author: Tel.: 734-945-6446; email: jbauer4@emich.edu
Manuscript submitted June 28, 2016; accepted March 25, 2017.
doi; 10.17706/ijcce.2017.6.2.111-126

Abstract: Cloud computing is a disruptive technology providing the occasion for change in the socio-technical structures of IT capacity management, but what are the latent effects of this? With interviews from ten case study organizations, this qualitative research generates a model that describes a spectrum of IT capacity-management structures, from classic to cloud, and describes the patterned differences that were discovered. Among the ten organizations studied, the latent, or unintended consequences of IT capacity-management trying to "stay relevant" during a transition to cloud computing adoption appears to lead to its own obsolescence. This analysis can be used as a platform for more targeted hypothesis testing to provide evidence for or against the generalization and external validity of this exploratory research.

Key words: IT capacity management, cloud computing, roles, processes, socio-technical structures.

1. Introduction

Cloud computing is a disruptive technology providing the occasion for change in the socio-technical structures of IT capacity management [1], [2]. What are the latent effects of cloud computing induced changes on the specific socio-technical structures of IT capacity-management? There are many recommendations and best practices on how to implement IT capacity-management processes [3]–[7]. However, there is a paucity of literature on the topic of how IT capacity-management processes are actually operated out in the field and because of that a qualitative approach was taken with ten case studies in order to start providing evidence for answering the research question. From the interview data, this research generates a model that describes the spectrum of IT capacity-management structures, from classic to cloud, and describes the patterned differences that were discovered. From this model it is possible to describe the latent effects cloud computing has had on IT capacity-management for the ten organizations that were studied. This analysis can then be used as a platform for more targeted hypothesis testing to provide evidence for or against the generalization and external validity of this exploratory research.

1.1. IT Capacity Management

The IT capacity-management field has many published frameworks to choose from. The IT Infrastructure Library (ITIL) is the most popular framework among IT service-management best practices with a 28 percent adoption rate in U.S. companies that are implementing IT service management [8]. Out of the six volumes published for defining ITIL, IT capacity-management is defined in the Service Design publication with about 20 pages of text [6].

At a high level, the ongoing activities of capacity-management follow a Deming-style iterative cycle of monitoring, analysis, tuning, and implementing [6]. ITIL recommends that operating systems, hardware configurations, and applications should be included in monitoring, which should also include information like utilization, transaction information, or response times [6].

In addition to ITIL, there is also the Capability Maturity Model Integration (CMMI) model collection for services, which describes capacity and availability management together as being responsible for ensuring effective performance and effective resource usage in support of service requirements [9]. Even enterprise architecture frameworks, such as the Open Group Architecture Framework (TOGAF), mention capacity-management from a change management and architecture perspective [10]. All of these frameworks provide advice for implementing IT capacity-management processes, but do not offer observations on how organizations are performing it in the field.

The Computer Measurement Group (CMG) is a vibrant international community of IT capacity-management practitioners. Founded in 1975, it aims to advance the field of IT capacity and performance management [11]. The CMG holds conferences, has a trade journal (Measure IT), and publishes a peer-reviewed journal called the Journal of Computer Resource Management [12]. A survey of the literature from the Journal of Computer Resource Management from 2000 to 2013 shows no articles observing or describing IT capacity-management processes as they are found in practice in organizations.

Some authors have interpreted and re-interpreted the ITIL framework as it pertains to IT capacity-management and how to implement it [5], [13]. For example, Lutz, Boucher, and Roustant [14] extend the ITIL framework for IT capacity-management to include better decision-making aids through modelling and monitoring activities. The authors identified two specific challenges they want to address: system complexity and business activities being isolated from capacity planning [14].

1.2. Cloud Computing

In simple terms, cloud computing delivers computing services to users at the time, location, and quantity they want, and at a cost that is based only on what is used [2], [15], [16]. It provides ubiquitous, convenient, and on-demand access to a shared pool of massively scaled resources that can be rapidly provisioned [2], [17].

It is called cloud computing, in part, because it abstracts out a layer of underlying infrastructure from the user of the service [16, p. 210], [17, p. 9]. Cloud computing services are commonly bound to the architecture layer they abstract, such as infrastructure, platform, or software [2, p. 8], [18, pp. 2–3]. Infrastructure as a service (IaaS) is where the hardware layer of the computing infrastructure is abstracted and provided as a service, like Amazon Web Services or Rack Space [2, p. 8]. A user of IaaS still has to install and manage an operating system and software. Platform as a service (PaaS) builds on IaaS by additionally abstracting out the operating system level of the architecture layer and providing that as a service, like Google App Engine or Windows Azure [2, p. 8]. Software as a service (SaaS) builds even further upon PaaS and abstracts out the software architecture layer and provides that as a service, like Google Docs or MS Office 365 [2, p. 8].

National Institute of Standards and Technology (NIST) provides a definition of cloud computing that is cited and re-used by others [2], [15]–[17]. It describes cloud computing as having the five essential characteristics of “on-demand self-service,” “broad network access,” “resource pooling,” “rapid elasticity,” and “measured service” [18, p. 2]. You can hear the definitions of cloud computing that opened this section as echoes of the NIST definition. It’s worth repeating: cloud computing delivers computing services to users at the time, location, and quantity they want, and at a cost that is based only on what is used [2, p. 7], [15, p. 1], [16, p. 211]. That covers the NIST defined characteristics of “on-demand self-service” (when I want), “broad network access” (where I want), and “measured service” (quantity and cost I want). Then, the other

part of the definition covers the rest: It provides ubiquitous, convenient, and on-demand access to a shared pool of massively scaled resources that can be rapidly provisioned [2, p. 6], [17, p. 19]. "Resource pooling" and "rapid elasticity" are covered by this part of the definition. It's worth noting that the concept of elasticity doesn't just apply to being able to quickly order more stuff. Elasticity is also the ability of the provider to dynamically add or remove resources within the abstracted architecture layer without the other layers being affected [16, p. 212].

NIST and others make a distinction between private and public cloud services [15, pp. 62–70], [17, p. 21], [18, p. 3]. In public cloud, the computing services are offered to the general public, and the infrastructure is on the premises of the cloud-service provider [17, p. 21], [18, p. 3]. With a private cloud, the service is just for one organization and is usually thought of as being internal to one organization [17, p. 21], [18, p. 3]. In a private cloud setting, the infrastructure is provisioned just for one organization [18, p. 3].

2. Research Design

The objective of the research is to provide evidence for answering the question of what latent effects cloud computing has on the socio-technical structures of IT capacity-management. To do that it generates a model that describes the current state of structures as operated at ten organizations, not how they were planned to be operated. To do this, semi-structured formal interviews with individual IT capacity-management practitioners were utilized

The subjects consist of twelve IT capacity-management practitioners. As will be seen in the analysis, not all of the subjects identify themselves as a capacity manager, but they all do perform activities consistent with IT capacity-management. Subject selection began by soliciting participation from practitioners who are members of the Computer Measurement Group (CMG), a community of IT capacity-management practitioners [11]. In order to get more participation from practitioners of capacity management in organizations with higher degrees of cloud adoption, the researcher made solicitations via LinkedIn.com professional groups. Twelve subjects from ten organizations were interviewed altogether. The organizations were from varying locations and of varying sizes. Six industries were represented: Insurance, Financial Services, Consulting, Telecommunications, Software, and Higher Education. The table below summarizes the organizational demographics for each organization being studied.

Table 1. Organizational Demographics Summary

Organization	Industry	Years doing Capacity-management	Number of employees	Location
Organization 01	Insurance	10	2,000	Midwest, USA
Organization 02	Financial Services	9	20,000	Western, USA
Organization 03	Consulting	9	1	Southwest, USA
Organization 04	Consulting	4	250	Eastern, USA
Organization 05	Telecommunications	2	40,000	Brazil
Organization 06	Higher Education	10	5,600	Midwest, USA
Organization 07	Telecommunications	20	100,000	UK
Organization 08	Software	3	30	Midwest, USA
Organization 09	Software	4	30	Midwest, USA
Organization 10	Higher Education	7	2,000	Midwest, USA

Except for Organization 06, each organization had one interviewee with one interview session. Organization 06 had three interviewees who were interviewed during three separate interview sessions. The table below summarizes the demographics of the individual interviewees.

Table 2. Interviewee Demographics

Person	Sex	Job Title	Percent Appointment	Highest Education	Years doing Capacity-management	Years with organization
Person 01	Male	Capacity Planner	50	Bachelor's	38	10
Person 02	Male	Enterprise IT Architect	0	Bachelor's	15	9
Person 03	Male	Chief Technical Officer	100	Ph.D.	25	9
Person 04	Female	Global Practice Principle	100	Bachelor's	25	4
Person 05	Male	Capacity Analyst Senior	100	Unassigned	Unassigned	5
Person 06	Male	Project Manager	40	Bachelor's	5	18
Person 07	Male	Capacity Manager	100	Bachelor's	10	40
Person 08	Male	Enterprise Infrastructure Architect	100	Master's	7	Unassigned
Person 09	Male	Chief Systems Engineer	100	Bachelor's	5	Unassigned
Person 10	Male	Director of IT and Software Development	100	Master's	4	4
Person 11	Male	Manager of Core Services	100	Bachelor's	20	5
Person 12	Male	Systems Administrator Senior	100	Bachelor's	17	7

3. Data Collection

The semi-structured interviews were guided by questions designed and selected in advance (see Table 3). In order to improve reliability, the interview questions were reviewed by a practitioner of IT capacity-management for clarity and understandability. If an in-person interview was not possible, a telephone-based interview was conducted. Three interviews were performed in person (for organizations eight, nine, and ten), while the remaining nine were held over the phone. Subjects were allowed to answer the questions however they felt and were not guided in their answers through feedback from the researcher. Audio recordings of the interviews were transcribed and loaded into Nvivo for analysis.

Table 3. Interview Questions

Primary Question	Possible Follow-up Question
In which field or industry is your organization?	
How many employees are at the organization?	
How long has your organization been doing IT capacity-management?	
Where is your organization located?	
Is there a framework or best practice that is being followed?	
What is your job title?	
Is capacity-management a full time roll for you?	
What is the highest level of education you have completed?	
Have you had any formal IT capacity-management training?	If so, what?
How long have you been practicing IT capacity-management?	
How long have you been with this organization?	
What is your role in capacity-management at your organization?	
Are there any other roles associated with capacity-management at your organization?	
Anyone else doing IT capacity-management with you?	
What is IT capacity-management supposed to do in your organization?	What needs does it fulfill?
How do you know when IT capacity-management is successful?	Who defines success?
What activities are carried out when performing capacity-management?	
What are the inputs of capacity-management?	Who does it come from?
What is the output of capacity-management?	Who receives these outputs?
What needs to be in place in order for capacity-management to function properly in your organization?	Example of when it went well? Or Bad?
Are there cases where people go outside of the capacity-management process?	If so, what is it they are doing?
Have there been any unintended consequences associated with doing IT capacity-management at your organization?	Have you noticed any positive or negative effects that weren't intended?
Anything else you think I should know about?	

4. Data Analysis

Because they were fielding so many questions about how they had approached the research for Awareness of Dying, Glaser and Strauss published their grounded theory approach in Discovery of Grounded Theory [22, p. 8]. Charmaz [23] built upon grounded theory with constructivist grounded theory, which acknowledges the reality that a researcher will come to a topic with some past history and biases but must manage them rather than ignore them. The researcher is coming to this topic with two decades of IT experience and will neither recuse himself from conducting such research nor pretend he is unaware of such past experiences. Instead, Charmaz' specific constructivist flavor of grounded theory will be used over Glaser's classic grounded theory wherever the two diverge.

Following a grounded theory approach, the data analysis consisted of three basic steps: open coding, selective coding, and development of a theory [22], [24, p. 143]. In this case, the "theory" is a descriptive model. During the first phase of coding one should compare incident to incident (in this case that would be interview to interview) instead of within a single incident [22, p. 39]. The goal is to find patterns within the entire body of data, not just a single source.

Selective coding occurs after open coding has hit category saturation [22, p. 75]. Category saturation occurs when the compared incidents can be interchangeable for the same concept, and the introduction of new incidents does not yield any new categories or concepts [22, p. 40]. Selective coding is used to guide further data collection and future coding so as to focus on only to the core variables of concern to the theory under development [22, p. 75]. See Table 8 for the list of codes used in this research. Tables 4, 5, 6, and 7 show a summary of the code references across the organizations studied.

Table 4. Organization References by Role Codes

Organization	Customers	Upper management	Finance or accounting	Other teams within IT
Organization 01	1	3	1	0
Organization 02	3	0	0	1
Organization 03	7	1	0	0
Organization 04	3	0	0	1
Organization 05	0	2	0	1
Organization 06	1	2	0	4
Organization 07	1	0	2	4
Organization 08	1	0	1	0
Organization 09	2	0	0	0
Organization 10	1	0	0	0

Table 5. Organization References by Input Codes

Organization	Business Projections	Cost	System Configuration	System Performance Metrics
Organization 01	4	2	0	0
Organization 02	1	0	1	5
Organization 03	0	0	2	2
Organization 04	0	1	2	1
Organization 05	3	0	1	1
Organization 06	0	4	2	1
Organization 07	0	0	0	1
Organization 08	0	1	0	0
Organization 09	1	2	0	0
Organization 10	0	0	0	1

Table 6. Organization References by Process Activity Codes

Organization	Delta analysis	Measuring and Monitoring	Modeling or forecasting	Moving load or modifying demand	Performance Tuning	Specify new system configuration
Organization 01	2	0	1	2	3	0
Organization 02	0	6	2	0	2	0
Organization 03	2	1	5	1	2	1
Organization 04	3	1	2	2	1	0
Organization 05	1	0	2	0	0	1
Organization 06	1	1	4	2	0	1
Organization 07	1	3	1	0	0	2
Organization 08	1	3	1	0	1	2
Organization 09	2	0	1	0	0	0
Organization 10	0	1	1	0	0	2

Table 7. Organization References by Output Codes

Organization	Forecasts	New System Configuration	Capacity recommendations	Reports	System Performance Metrics
Organization 01	2	0	0	1	0
Organization 02	0	4	1	1	5
Organization 03	0	0	0	3	2
Organization 04	1	0	3	0	1
Organization 05	1	1	0	1	1
Organization 06	1	2	3	0	1
Organization 07	0	2	2	0	1
Organization 08	1	5	0	0	0
Organization 09	0	2	0	0	0
Organization 10	2	1	3	0	1

Finally, during the development of a theory, a model or explanation is proposed that is grounded in the data that were collected [24, p. 143]. In this research a descriptive model is generated. The model that was developed is based on the codes from the previous steps [23, p. 63].

The analysis for this research began after the second interview with open coding and continued after each interview. The transcripts were read line-by-line and at least one category was created or reused for each sentence. This generated a large number of categories, but over time it was easy to see which categories were truly patterns and which were not. While coding for the fourth and fifth interview transcripts, it was clear that category saturation had been hit. No new categories were being created and only existing categories were being used. From there, the line-by-line categories were analyzed for whether they represented a true pattern across subjects.

To improve reliability, transcripts were checked for any obvious errors before coding began [25, p. 190]. Nvivo software was utilized to aid in tracking and managing the coding, categorizing, and relationship building during the analysis. The interview transcripts were put into the software, and any themes, codes, or categories were applied directly to the text within the software.

To minimize code drift the transcript portions that matched a code were compared with each other to make sure the meaning of the code did not drift during the coding process [25, p. 190]. This was handled operationally by opening a coding node within NVivo, which presents a view of all the associated text that was coded in one window. A code dictionary was also used during coding to assist with consistent code application (seeTable 8). The definitions of these codes were put in the node property notes within the NVivo software so that they were easy to reference during coding activities.

Table 8. Code Dictionary

Theme	Category label	Description
Activities	Cost or Budget Analysis	Financial calculations.
	Delta analysis	Comparing one variable against another. For example, amount of resources on hand vs. amount of resources consumed.
	Measuring and Monitoring	Collecting, filtering, sorting information.
	Modeling or forecasting	The creation of mathematical models to describe or forecast.
	Moving load or modifying demand	Changing where jobs or tasks (work) is performed. Such as storing files in a different location. Or, changing the behavior of user activity. For example, setting specific hours users can access a certain system.
	Performance Tuning	Finding areas where the technical architecture or configuration can be changed in order to run more efficiently or for work to complete more quickly (or both). For example, accomplishing the same tasks with fewer system resources or servers.
	Specify new system configuration	The act of creating specifications for new technical system configurations (add/remove/modify).
	Training or Mentoring	Supplying others with knowledge, wisdom, or advice. Sometimes a single or limited transaction, other times an ongoing relationship.
Inputs	Business Projections	The projections, plans, or goals from the business or non-IT area of the organization.
	Cost	Information about the cost of resources or services.
	System Configuration	The current technical system attributes that are owned or managed by the organization. For example, how much memory or CPU is installed in a given server.
	System Performance Metrics	Information about the technical performance of systems. Things like CPU, memory, network, etc. Not how much is installed (that's system configuration) or how it is configured, but how much is being used.
Outputs	Forecasts	Predictions about the future.
	New System Configuration	Added (including buying), removed, modified technical system configurations. For example, added or removed memory or CPU.
	Capacity recommendations	Suggestions or recommendations to a different role for action.
	Reports	A document (physical or virtual) given to or accessed by another role. Could contain other outputs, like recommendations or forecasts, for example.
Related roles	Customers	The people who purchase or use the services the organization offers. In the case of consultants, the customer is the organization that hired them to provide IT capacity-management as a service.
	Upper management	Represents members of upper management all the way up to the CxO level (or equivalent).
	Finance or accounting	The role that manages the money and budget for the organization. Sometimes they are also the purchasing authority.
	Other teams within IT	Teams within the area of the organization that handles IT, but not the team that has the IT capacity-management role in it. It most commonly includes teams such as analysts, architects, data center operations, programmers, application designers, or other operational IT teams.

Reliability was also improved through the use of a panel of judges, who reviewed the application of the codes to the transcripts. Intercoder agreement means that the judges agreed that the codes match the transcript sections to which they have been applied [25, p. 191]. Intercoder agreement of 80 percent or greater indicates good qualitative reliability [25, p. 191]. The panel was made up of three academic peers familiar with qualitative methodologies and intercoder agreement was found to be 87%.

A panel of judges, similar to the one used for intercoder reliability, was used to review validity. This is

similar to the peer debriefing that Creswell recommends, and it focuses on whether the codes themselves make sense and represent what they purport to represent [25, p. 192]. As with the intercoder reliability panel, 80 percent agreement indicates good validity [25, p. 191]. The members of this panel were mixed, with some academic peers with sociology experience and some with IT backgrounds. Agreement for the panel was found to be 100%.

5. Model for Spectrum of IT Capacity-Management Structures

From the very first interviews, the topic of cloud computing came up over and over, even from subjects in organizations not using cloud. While performing open coding it became clear that there was some sort of spectrum of IT capacity-management structures. It seemed that on one side was a sort of classic IT capacity-management structure, which emerged from the old mainframe days when computers were extremely expensive. Then, on the other side was a more emerging cloud IT capacity-management structure, where capacity is seen as virtually unlimited. In this structure, not all capacity-management activities were performed by people with a job title that contained the words "capacity-management." It seems this is especially true the closer to the cloud side of the spectrum you go. The cloud appears to be a disruptive technology that is providing an opportunity for structural change, where the roles, actors, and structure are all in flux [1]. Some of the organizations seemed to fit one of the extremes or another of a spectrum, but others were something of a blend between the two extremes. Therefore, a spectrum of IT capacity-management structures is presented (see Figure 1).

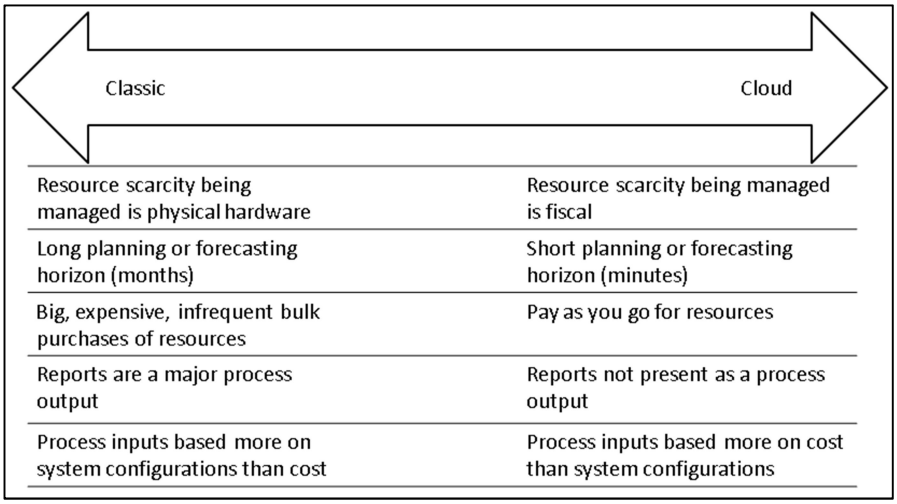

Fig. 1. Summary of spectrum of IT capacity-management structures.

While interviewing subjects about the field of IT capacity-management, there was a general feeling among many subjects that, as a profession, it is in trouble or at least in the midst of some big changes. Sometimes these sentiments would come up while answering questions about their role and they would describe how things have changed for them and their bleak sounding feelings about the future. Other times they would talk about it at the end of the interview, where they were asked if there was anything else the researcher should know about. For example, Person 3 confided that there are not too many people in the field of IT Capacity-management.

> **Person 3:** "I was talking about it and I always try to talk to people in the same field and unfortunately to be honest with you there are not that many."

The perception that classic IT capacity-management is waning is reinforced by observations that classic

IT capacity-management roles are disappearing from organizations. As an example, while asking Person 7 about the roles associated with IT capacity-management in their organization they described how they work on managing the capacity for their organization's private cloud (what they call an "enterprise cloud") and took a moment to reflect on recent history:

> **Researcher:** "Are there others doing IT capacity-management for things other than the enterprise cloud?"
> **Person 7:** "I'm just thinking… there used to be. They closed a lot of the departments. They were thinking they don't need quite as much -- as many people doing the job as they used to."

This general sense of the profession as one that is shrinking leads to another trait found among the classic IT Capacity Managers. Something of an urge for professional self-preservation. If the role of IT capacity manager is diminishing, the actors of that role could feel they need to re-assert their relevance in order to preserve their professional role. This can be seen, for example, when Person 01 feels that it needs to be said that IT capacity-management is still relevant as a discipline.

> **Person 1:** "Yes, it's things like your [research] hopefully can bring about some of the need for saying [IT] capacity [management] is still very relevant today."

The shrinking professional role and the perceived need to keep the practice of IT capacity-management relevant stem, in part, from a cost to benefit ratio that has flipped over the last couple of decades [2, pp. 10 – 16]. At the end of the interview, while asking if there was anything else the researcher should know about, Person 2 offered an historical perspective of the cost benefit ratio.

> **Person 2:** "So, the nature of capacity planning is just fundamentally changing because the ratio between people costs to hardware costs has changed so fundamentally. It's inverted over the last 20 years."

The cost ratio has inverted because of the declining cost of computing, but also because there is a perceived change in the nature of the operational management of computing resources [2, pp. 10 – 16]. Virtualization and private-cloud computing change the way computing resources can be managed and the cost in managing them, too [2]. It is at this point we begin to step from the left part of the spectrum and toward the right side; away from classic IT capacity-management and toward cloud. After giving their historical perspective on the cost and benefit ratio, the researcher asked Person 2 what he/she thought the future held for IT capacity-management, and you can see them wrestle with the common sense impact of the inversion of the classic capacity-management cost ratio.

> **Person 2: "**With the virtualized environments it's like, dude, why don't you just grow? You don't need a capacity planner, you need a loading dock supervisor -- you just keep shoving stuff to the floor, and you let the system auto-balance. And you're able to do that essentially. There are some very good products that allow that kind of stuff to occur on the floor. Before it was just -- it'll tell you automatically where it's recommending you add machines but increasingly there are products that say I'm going to do it for you -- just give me all the resources I'll take care of it -- don't you worry yourself about it."

Person 2 is talking about automated virtualization resource management, where the management software automatically moves the workloads around on the resources for you. Because of that the unit of management shifts away from being the individual servers to the overall capacity of the data center. He says you just need a loading dock supervisor because at that point the resource management activities that are

left to perform are really just the logistics of buying and installing new hardware when the management software tells you it needs more (" . . . shoving stuff to the floor. . . " (Person 2)). Person 2 was talking about the changes in operational management when virtualization is involved with locally owned resources. It begins to introduce a lot of the characteristics and attributes of cloud, but still remains on premise and on locally owned hardware. On the spectrum of IT capacity-management structures virtualization would fall somewhere near the middle. Because of virtualization and the ability to abstract physical resources, it takes on many similar characteristics as cloud, but still requires physical resources hosted on-site.

On the far right side of the spectrum is the cloud IT capacity-management structure, where there are no local physical resources purchased and capacity is so readily available that it cast the illusion of being infinite. But if capacity is infinite, then what is the resource that is managed? While Person 8 talked about his/her process activities and outputs as he/she shifted toward using cloud solutions, the researcher asked a follow-up question around a shift in what kind of resources Person 8 was managing when he/she started moving to the cloud.

> **Researcher:** "The resource scarcity has changed?"
> **Person 8:** "Yes, absolutely."
> **Researcher:** "So what's the new resource scarcity?"
> **Person 8:** "The new scarcity?"
> **Researcher:** "Yes."
> **Person 8:** "I guess it's purely money that you're optimizing now and, you know, in somewhat real time."

Because there are no physical resources to purchase and manage as a scarce resource and you pay only for what performance is used, the cloud IT capacity-management structure is more focused on financial aspects of vendor-relationship management and contract management.

> **Person 8:** "When you're dealing with the public cloud issue, capacity itself isn't really any longer an issue from the perspective of planning to purchase hardware. You know, how large a tape robot[1] are we going to buy? How many NAS[2] devices are we going to put in, how we're going to interconnect them. Those kinds of things that we used to think about become a little less relevant and it's more of -- so, how do we analyze say, for instance, what's the differences between Amazon S3 and Amazon reduced redundancy S3 and Glacier and do we trust having data in Amazon in one region, two regions? Do we want multiple cloud vendors? And then essentially it's kind of a cost optimization, particularly in the compute kind of world and the network world."

As organizations transition into the use of cloud, other existing professional roles do not simply disappear. Many organizations already have people working on legal and financial activities. If IT capacity managers are going to manage the financial and contractual aspects of capacity (since the physical aspects are disappearing), then they could find themselves in competition with people in their organization who, at least nominally, already do that kind of work. This latent effect of cloud computing generates a struggle that is even observed by the subjects. Person 11 offered his/her view of the future at the end of the interview:

> **Person 11: "**Kind of like I said before, capacity-management in the cloud really isn't what in the traditional sense you'd call capacity-management. It's more management of dollars and cents and efficiency management more than anything else. Capacity-management in general, I think, starts to get affected from a skill set or a way of working. As cloud and or colocation or co-hosted data centers crop up, you kind of leave the aspect behind of buying capacity as a large pool of resource

[1]Tape robots are used to physically exchange specialized cassette tapes from a storage area to a machine that reads or writes to them. The cassettes are used as inexpensive massive storage for backing up data.
[2]Network Area Storage, a type of hardware storage device that stores data.

> and then trying to figure out how to apply your services or objects that are going to consume the resources to it to gain the maximum value out of the investment. The cloud really changes that fundamentally because you kind of forget doing that because the fundamental difference in the cloud is I don't pay for the big chunk up front and then try to use it. I just try to manage and use what it is I actually need to efficiently deliver the service."

Person 11 is getting at the resultant shift of what is managed when you move from the classic IT capacity-management structure toward the cloud IT capacity-management structure. Since there is no local hardware to purchase in bulk, the scarcity that is being managed changes. The 'thing" that is measured changes. The capacity being managed is fiscal in nature. Because there is already a role in organizations for managing "things" of a fiscal nature (accounting or finance), it sets up a conflict within the social patterns that create the structure. The resolution to this conflict, among the organizations interviewed, is for IT capacity managers to become more closely linked to the data center operations (during the middle, or transitional part of the spectrum) and then for the formal role to disappear (in the cloud end of the spectrum). While Person 7 was talking about the outputs of their process and the roles that use them, it struck the researcher that a social shift was occurring for them as they moved toward cloud solutions.

> **Researcher: "**Do you feel kinda socially closer to the application teams or the data center teams?"
> **Person 7: "**Probably now to the data center teams. They're trying to close in the silo so we don't talk to the applications. I don't know if that's strategy or side effect, but I used to get very close to the applications, but with the virtualization we're a lot closer to the data center -- the configuration teams. How many blades you need to put in a rack and that sort of thing. So yes, it used to be we were closer to the apps; now we're closer to the data center."

When taken together, the change in which resource is being managed, along with which professional roles that are managing them, a consistent view from the subjects of the far right position of the spectrum is one where the professional role of IT capacity-management is anachronistic and unnecessary for the most common organizational needs. At the end of the interview, Person 9 offered his view on the professional role of IT capacity-management and where it is likely to go.

> **Person 9: "**The fact of the matter is there are still people who do capacity planning, and it's a really cool job because they do it at places like Amazon, and Google, and Facebook. And they're dealing with a scale that mom and pop shops who are running a thousand servers will never see. When you start talking about hundreds of thousands or millions of hosts then you have really interesting problems. So, you have fewer people who get hired but are at the top of their game. The time of that being a title that people have is fading rapidly."

This view is one where classic IT capacity-management would remain a professional role only in a select few environments where the financial ratio still makes sense. These are the very large computing environments like Amazon, Facebook, Google, and so on. This is where the cost of resources (because of the volume of them) is still quite higher than the cost of hiring people to manage their capacity. For the majority of the remaining organizations, where the financial ratio doesn't make sense, the activities performed by a capacity manager are distributed to other existing roles. While answering a question about roles associated with process outputs, Person 8, from an organization that doesn't have a formal capacity manager (Organization 6, which is doing both cloud and local virtualization and placed in the middle of the spectrum), describes the various roles involved with managing capacity for them.

> **Person 8:** "Our virtualization group, which manages our internal virtual environment, also looks at

> a lot of the trending information in our cloud virtual environment. Our business office[3] receives the bills, and those two groups go through them looking for the coarse grain kind of 'where can we optimize' things. And they'll go back to the services involved."

Nowhere in that brief process summary is the professional role of IT capacity-management mentioned. Yet, activities around management of resources are taking place. In this kind of environment it doesn't make sense to hire a dedicated IT capacity manager. At the end of the interview with Person 9, during informal conversation, the topic of the cost to benefit ratio came up:

> **Person 9:** "Why would you hire someone for something one of us could do in like 5 hours per month?"

The financial cost ratio and shift in what resources are managed delineate the far ends of the spectrum of IT capacity-management structures, but they also contain an undercurrent of deterministic progression, where the increasing adoption of cloud computing leads to the inevitable demise of the classic IT capacity-management profession as a role. Some were less nuanced about this than others were.

> **Person 9: "**Very few people make carriage axles as well. Sometimes the old ways were bad."
> **Person 10: "**I like to think that as technology improves it actually improves."

While walking through the spectrum, the differences in the structures become evident. You can see in Figure 2 that the organizations studied were somewhat balanced across the spectrum. Four fall within classic capacity-management, three in cloud, and three were somewhere in between classic and cloud. An organization could fall in between by having both local resources for some services but also using cloud (public or private) for other services.

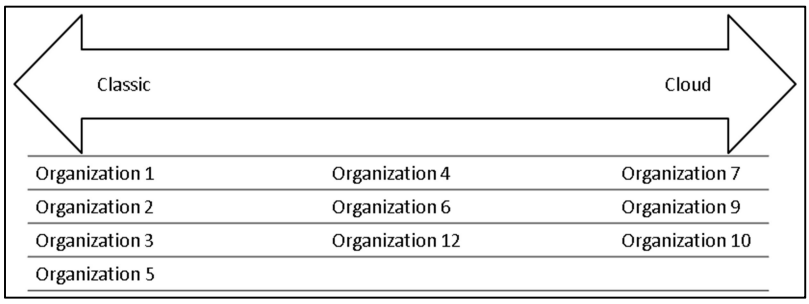

Fig. 2. Interviewed organizations plotted along the IT capacity-management structures spectrum.

5.1. Latent Effects of Cloud Computing on IT Capacity Management

Cloud computing appears to be playing the role of a disruptive technology that is providing the occasion for changes in IT capacity-management structure [1], [2]. While this study did not follow a single organization in a longitudinal fashion to observe its structural transformation from the classic side of the spectrum to the cloud side of the spectrum, several organizations which were on one side or another, or in transition, were studied. By piecing together the patterns of roles and their relationships a tentative model of change across the spectrum can be constructed (See Figure 3). Figure 3 shows the observed relationships between roles within IT capacity-management structures along the IT capacity-management structures spectrum. This observational model shows the differences in the roles that are associated with operating IT capacity-management processes within the studied organizations as one moves focus from classic to cloud. This is not a model of technical or even organizational structure. It also only shows the social roles related

[3] The business office handles finance, legal, and accounting activities.

to the operation of the IT capacity-management processes as described by the subjects. This model is based only on the organizations studied in this research, so future research would be needed to provide evidence for external validity. Ideally, an organization would be studied longitudinally from the classic structure all the way through its transition into a cloud structure.

In the classic structure, the IT capacity manager role is something of a mediator between upper management and the rest of the IT roles. The "application development role," represents those who have a role in application development, user experience design, business analysis, and so on; then "data center and systems administrator roles" is the role for those who work on the physical hardware up to the operating system. In this structure the capacity manager commonly works with upper management to get business forecasts and requirements, then works with application development roles, data center and systems administrator roles to monitor and measure the systems. Then the capacity manager generates a report with capacity recommendations for upper management to consider. In this sense, the capacity manager plays something of a translator role between upper management and the other IT roles. In social network terms, the capacity manager has the power of "centrality" in this mediator role, since all the other roles pass through it [26].

As you move toward the cloud structure and enter the transitional area a curious thing seems to occur with the capacity-management role in the transition. The capacity-manager role gives up their power of centrality and position as mediator and bonds themselves closely with the data center. Some evidence was provided to suggest that this move was a bid to retain functional relevance in the organization. It could also be a response to the inversion of the cost ratio between hardware and staff, since an entire data center certainly cost far more than any individual pieces of IT infrastructure. A full time employee is more justified at the data center scale than the scale of individual servers that may only cost a few thousand dollars. Additionally, in order to continue using the skills from the classic structure, the capacity manager could apply their function to the data center with little required change in their skills or function.

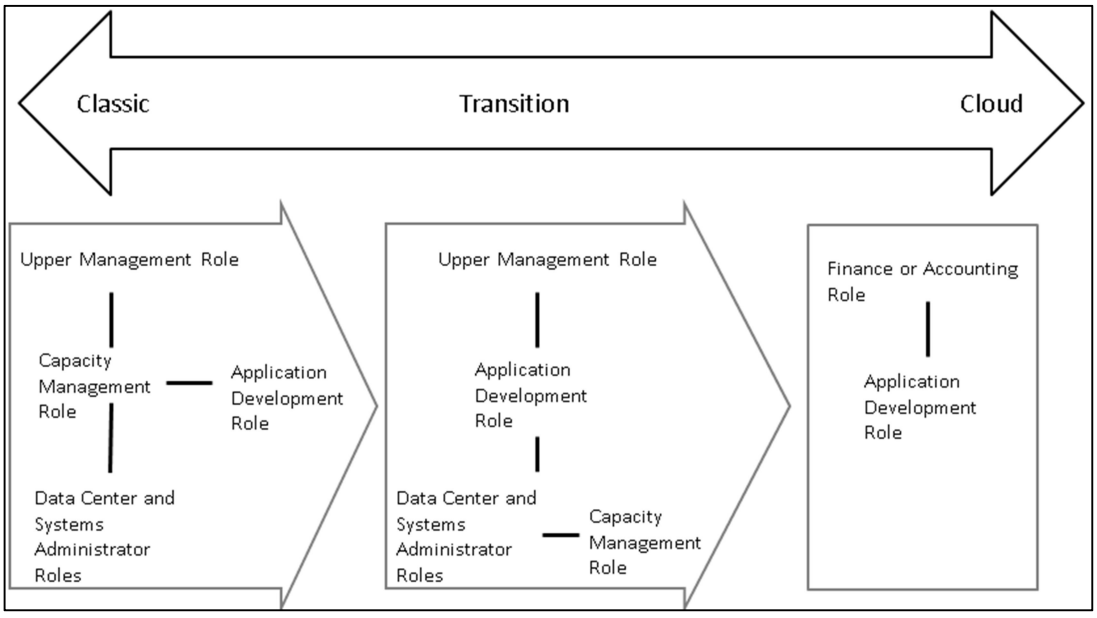

Fig. 3. The observed role relationships in IT capacity-management processes.

As virtualization (often a localized precursor to cloud computing) comes in, its resource management is such that the application and operating system level capacity-management is handled automatically by the virtualization software. Therefore, a slight shift in what the capacity manager manages occurs. They end up

managing the capacity of the data centers, rather than myriads of individual applications or servers. This shift bonds them closely, socially, with the data center and heavily de-emphasizes the relationship with the application development and upper management roles. An example of this was described quite succinctly by Person 7 while reflecting on their organization's transition:

> **Researcher: "**Do you feel kind of socially closer to the application teams or the data-center teams?"
> **Person 7: "**Probably now to the data-center teams. They're trying to close in the silo so we don't talk to the [application developer roles]. I don't know if that's strategy or side effect, but I used to get very close to the [application developer roles], but with the virtualization we're a lot closer to the data center -- the configuration teams. How many blades you need to put in a rack and that sort of thing. So yes, it used to be we were closer to the apps, now we're closer to the data center."

Of course, one of the defining traits of cloud computing is that you do not have to bear the cost or risk of owning a data center. Part of the cost and risk of owning and operating a data center is in staff. So, as you move over to looking at the pure cloud structure on the spectrum, all the social roles related with data center management are absent because that work is no longer done at the organization. The cloud abstracts out the physical hardware and even the operating system if you want. Therefore, a data center and all the roles associated with it are no longer required. This is a manifest function of cloud computing and is sought out on purpose. However, the latent effect of this on IT capacity-management structures, since it bonded itself with the data center during the transitional phase, is that it disappears as a discrete role along with the data-center roles. As the research has shown, this does not mean that capacity-management activities are no longer performed within the organization, they just change slightly in nature and are performed by different roles (finance or accounting, and the application-development roles, for example). Also, those interviewed in the cloud structure had stronger relationships with accounting and financial roles than they did with upper management. This is likely because resource procurement is done more in an incremental and subscription-based model and no longer in the form of infrequent, massive capital expenditures that require senior-leadership approval.

Based on this tentative model, it appears that if an individual capacity manager wants to remain relevant as an organization moves from a classic structure to a pure cloud structure, he/she should find a way to bond himself more closely to the application development and finance or accounting roles and should probably look to divest themselves of the formal title of capacity manager. In other words, they should invest in application development and finance skills and change their primary role. This could prove difficult depending on the social background of the individual capacity manager. In the researcher's 20 years of experience in IT, he has observed something of a social stratification within the culture of IT where those who deal with the application layer have a higher social stature than those who physically stack servers and string cables in a data center. It is a social structure that is often reflected in boundaries of teams in the organizational structure, which, itself, is often a reflection of the structure of IT itself. It may prove difficult for a capacity manager who comes from more of a data center or server hardware background to bond themselves more closely with application layer people because of this social stratification. Whether social stratification within IT culture is a factor in capacity-manager longevity while an organization moves to a cloud structure is another topic for future research.

6. Conclusion

Evidence has been provided to illustrate the latent effect cloud computing is having on the structures of IT capacity-management. Among the ten organizations studied, the latent, or unintended consequences of IT capacity-management trying to "stay relevant" during a transition to cloud appears to lead to its own

obsolescence. The analysis here is presented so that practitioners may begin to study it to see if this model can be generalized beyond the ten organizations studied and, if so, to then take steps to either adapt the field of IT capacity-management or attempt to reshape the path for the role as organizations transition to cloud structures. Future work should include research to establish evidence for or against the external validity of the model of role relationship structures summarized in Fig. 3.

References

[1] Barley, S. R. (1986). Technology as an occasion for structuring: Evidence from observations of CT scanners and the social order of radiology departments. *Administrative Science Quarterly*, *31*, 78–108.

[2] Kushida, K. E., Murray, J., & Zysman, J. (Mar, 2015). Cloud computing: From scarcity to abundance. *Journal of Industry, Competition and Trade*, *15(1),* 5–19.

[3] Augello, M. (2000). A Corporate-wide approach to capacity and performance management. *Journal of Computer Resource Management*, *97*, 2–5.

[4] Gunther, N. J. (2010). *Guerrilla Capacity Planning: A Tactical Approach to Planning for Highly Scalable Applications and Services*. Berlin: Springer-Verlag.

[5] Molloy, C. (Oct. 2003). Using ITIL best practices to create a capacity management process. *Measure IT*.

[6] Office of Government Commerce. (2011). *ITIL Service Design*, 2nd ed. Norwich, UK: The Stationery Office.

[7] Sheldrake, H. (2009). A framework for capacity planning. *Journal of Computer Resource Management*, *124*, 23–35.

[8] IT Governance Institute. (2011). *Global Status Report on the Governance of Enterprise It (GEIt)—2011*.

[9] CMMI Product Team. (Nov. 2010). CMMI for services, version 1.3. *Software Engineering Institute Carnegie Mellon*. Retrieved October 9, 2012, from http://www.sei.cmu.edu/reports/10tr034.pdf

[10] TOGAF. (2011). 16. Phase H: Architecture change management. *The Open Group*. Retrieved October 9, 2012, from http://pubs.opengroup.org/architecture/togaf9-doc/arch/chap16.html

[11] Computer Measurement Group, Inc., "About CMG," *Computer Measurement Group*, 2012. [Online]. Available: http://www.cmg.org/national/about-cmg.html. [Accessed: 08-Oct-2012].

[12] Computer Measurement Group, Inc., "CMG publications," *Computer Measurement Group*, 2012. [Online]. Available: http://www.cmg.org/national/publications.html. [Accessed: 08-Oct-2012].

[13] Grummit, A. (Sep. 2009). Knot-ITIL: How to solve the gordian knot of IT service management using the sword of best practice. *Measure IT*.

[14] Lutz, M., Boucher, X., & Roustant, O. "Methods and applications for IT capacity decisions: Bringing management frameworks into practice - ProQuest," *Journal of Decision Systems*, *22 (4)*, 332–355, 2013.

[15] Fehling, C., Leymann, F., Retter, R., Schupeck, W. & Arbitter, P. (2014). *Cloud Computing Patterns*. Vienna: Springer Vienna,.

[16] Kushida, K.E., Murray, J., & Zysman, J. "Diffusing the cloud: Cloud computing and implications for public policy," *Journal of Industry, Competition and Trade*, *11(3)*, 209–237, Sep, 2011.

[17] Beri, R. & Behal,V. "Cloud computing: A survey on cloud computing," *International Journal of Computer Applications*, *111(16),* 19–22, Feb, 2015.

[18] National Institute of Standards and Technology, "The NIST definition of cloud computing." U. S. Department of Commerce, Sep, 2011.

[19] Merton, R. & Sztompka, P. *On Social Structure and Science*, University of Chicago Press, 1996.

[20] Merton, R. *Social Theory and Social Structure*, 2nd ed. Glencoe, Ill.: The Free Press, 1957.

[21] Giddens, A. & Dallmayr, F.R. *Profiles and Critiques in Social Theory*. London: Macmillan, 1982.

[22] Glaser, B.G. *Basics of Grounded Theory Analysis*. Mill Valley, CA: Socioloty Press, 1992.

[23] Charmaz, K. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. London: SAGE Publications, 2006.
[24] P. D. Leedy, P. D. & Ormrod, J. E. *Practical Research: Planning and Design*, 9th ed. Boston: Pearson, 2010.
[25] Creswell, J. W. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, Third. Thousand Oaks: SAGE Publications, 2009.
[26] Freeman, L. "Centrality in social networks conceptual clarification." *Social Networks*, *1,* 215–239. 1979.

Joe Bauer is a research computing consultant at University of Michigan (Ann Arbor, MI) and adjunct faculty at Cleary University (Howell, MI). His research interests include the explication of new socio-technical structures related to the onset of the information age and the identification of opportunities for democratization of technology management. Dr. Bauer holds a doctorate degree in technology management from Eastern Michigan University.

Al Bellamy is a professor of technology management in the College of Technology at Eastern Michigan University in Ypsilanti, MI. His interests in applied research include organizational factors affecting the implementation of information technologies and leadership and emotional intelligence. He holds a doctorate degree in sociology from Purdue University.