PERSPECTIVES ON COMPUTING FOR SERVICE PROVIDERS OF INTELLIGENT ENVIRONMENTS

Vijay Masurkar

Sun Microsystems, Inc., Burlington, MA, USA

Abstract; Intelligent Environments (IE) have been of interest to many modern research enterprises, particularly in advanced countries. For example, a recurring theme emerging from the European Commission regarding the Information Communication Technology (ICT) research agenda for the next 5-10 years is the notion of “Ambient Intelligence” (AmI) representative of intelligent environments and advanced services. As societies adopt such services, demands will surge on their provider infrastructures, and the environments will emerge, take shape and expand. These IEs will possess certain characteristics that will likely drive innovative operational developments in terms of dynamic policies for operations and management, service pricing structures, and infrastructure fault management and diagnosing strategies. This paper focuses, in broad terms, on these critical IE provider’s infrastructure needs.

Keywords; Intelligent, ambient, environment, infrastructure, provisioning, utility, computing, service, pricing, policy, model.

1. INTRODUCTION

Intelligent devices are everywhere; yet, in the true sense, few intelligent environments exist. The ones which exist are confined within the walls of the research laboratories. In the near future, however, dependence on them will likely increase and hence will become more or less natural. In some parts of the world, wearable computing, where human activity is the central part of the user’s context, is likely to be an upcoming fashion. A recurring theme emerging from the European Commission regarding the Information Communication Technology (ICT) research agenda for the next 5-10 years is the notion of “Ambient Intelligence” (AmI) representative of intelligent

environments, IE. This concept describes an environment where *“people are surrounded by intelligent intuitive interfaces that are embedded in all kinds of objects and an environment that is capable of recognizing and responding to the presence of different individuals in a seamless, unobtrusive and often invisible way.”* In the context of this paper, we can assume providers of IE primarily fall into two primary categories. Infrastructure providers such as vendors of network equipment or web servers, and service providers that serve travel and other end user needs with web services or back-end applications. The underlying assumption is that the latter type rely on the former one which covers the lower levels of the ISO 7-layer network protocol hierarchy.

The prominent characteristics of an IE that stand out are size, mobility, heterogeneity, complexity, knowledge and security.

- 1) Size: In pervasive computing, there are millions of subjects and objects, - many of them interacting with each other.
- 2) Mobility: Mobility introduces more vulnerability than in a static world. There is the necessity to safeguard against the danger that the physical size of a mobile unit or its ability to be location independent can be exploited by miscreants.
- 3) Heterogeneity: One cannot assume closed, co-designed systems in the world of IE. The permutations and combinations of interactions are countless.
- 4) Complexity: Both hardware and software complexity increases and the dependability matrix can become overwhelming; yet with open interfaces, the complexity must be minimized.
- 5) Knowledge: Distribution of knowledge, coupled with cooperating entities sharing knowledge, logarithmically increases with inter-communications between subjects and objects.
- 6) Security: On demand, peer-to-peer networking between objects drives requirements for secure communication that will be policy based (defined for a set of participating object types) for the duration of the communication. This will have (mutual peer level) authentication and authorization.

These characteristics, in turn, require that providers of infrastructure and various end user services take into account the constantly changing service requests of data, audio and video at different scales from static and mobile users, as well as users interacting with each other. As it is typically understood, intelligent environments are inherently attributed to be social and collaborative spaces. Consequently, users willingly subscribe to services for their social or business needs. So, behind the scenes, when large scale environments of such kind are deployed, these interactions can result

in a huge number of requests that need to be translated into compute, storage or network bandwidth requirements at the infrastructure level of an “intelligent” service provider and delivered reliably and securely via subscribed services.

Grid computing is typically employed for such dynamic resources monitoring and allocation, and there are several vendors providing such solutions; N1 Grid SPS (2004) and I-Fabric (2002) from Sun, HP Utility Data Center (2001), ThinkDynamics (2003). There is also the initiative that promotes an open industry forum, OGSi Working Group (OGSi-WG, 2004), to review and refine the Grid Service Specification and others including OGSA-infrastructure-related technical specifications. Although this may form some basis for the architecture, for the IE user, there is a lacuna to be filled in terms of service provisioning where user’s needs based service usage can be tracked and measured, and the pricing structure is low risk.

2. ‘MARIA – ROAD WARRIOR’ SCENARIO

One scenario in the European Commission’s ISTAG committee report concerns Maria who has traveled ‘long-haul’ to a foreign country to attend a business meeting at which she will give a presentation (Ducatel et al, 2001). The only computing system she needs to take on her trip is her ‘P-Com’, worn on the wrist. The P-COM interacts with external systems as Maria travels internationally, and allows her to travel unhindered by identity checks (at the airport), car rental arrangements or city-center travel restrictions.

See *Figure 1* below, which depicts P-COM’s interfaces to the outside world. The device allows voice communications, both private calls and Maria’s verbal instructions to nearby intelligent systems, and can store preferences and download video, on demand, at her hotel room. Maria has a biosensor within her P-COM that helps her to keep track of her own health.

Figure 1. P-Com and Maria's Business Communication

3. IMPLICATIONS ON THE INFRASTRUCTURE ARCHTECTURE

A service provider in the above scenario may be a rental car agency or a hotel chain. Service requests could range from simple identity verification by a foreign country's immigration department to downloading of a real-time video on a hand-held device. User demands for services will be more or less random, unpredictable and will typically scale quickly up and down. Adaptive, highly available and secure computing abilities are needed. There is, of course, the traditional capacity planning model, but would it suffice?

There is the emerging utility computing (UC) model where services can be provisioned as per the needs of the consumer, and consumer's service costs can be tracked and measured in real time. In the utility computing (business) model, customers neither incur high fixed costs of purchasing system hardware and software, nor commit to long-term fixed price outsourcing contracts (Paleologo, 2004). Both are seen as threats to adoption in traditional computing. Instead, in the UC, they receive the service they need and pay for its usage just as they would pay for water or electricity.

4. USERS PAYING FOR SERVICE USAGE

Utility computing based services reduce the risk faced by the customer. This is because the costs to the customer are proportional to the volume of transactions performed during a certain service usage interval. Secondly, UC deployment comes with an advantage of economies of scale. Applying this to the IE context, as the number of services hosted onto the infrastructure increases, the hardware, software and (operational and management) labor costs grow sub-linear to the total volume of transactions. In turn, the infrastructure transfers the lower cost of running an end user service to the users. So, the critical questions that service providers face regarding pricing are: what to price, how to price and when to price. Should a single unit of service cost be fixed for a pre-determined duration? If it is designed to be variable, what specific stipulations must affect the price? Take the above specific example; should a content delivery service to Maria's P-COM include guarantees on performance such as maximum latency for the end user service provider or packet loss for the infrastructure provider?

For all the "*intelligent*" services to be delivered, payment is required from the user to the provider. When Maria downloads a video into her hotel room, that is '*on-demand*' service; notice that it's also portrayed as a location-independent service offering. The implications are significant. Her identity will have to be verified through P-COM. Is the authentication and policy database replicated in the foreign country she is visiting? Or, is there a remote query to her country of residence? Are the digital download, access and copyright laws complied with by the download-on-demand service? How is the payment method set up for the foreign money exchanges? How are the service objectives, service and usage policies, service level agreement (SLA) and payments established?

The actual mechanisms may vary depending on the type of service; however, the principles remains the same. Service providers of IE infrastructures (who are customers of the infrastructure vendors) and end users of services will primarily look for low risk and high availability, reliability, security and performance (via SLA). Pricing will be standardized as it will be public. Yet, the actual determination of the price per service needs to take into account some risk factors associated with UC; in particular, their relative short lifetimes and high initial investments such as for building the infrastructure, for example (Paliologo, 2004). Short lifetimes result from the fact that a user of the services (i.e. an end user) can switch from one service provider (or vendor) to another without much loss of investment on his or her part. Further, note that their adoption could be highly uncertain due to the pace of technological innovation in this area.

5. MEASUREMENTS FOR SLA COMPLIANCE

How can automatic measurements be done of the service level performance? This is particularly important from the perspective of the users so they can compare the actual performance of their subscribed service with the agreed level defined in the SLA with their service provider. SLAs are primarily business contracts between the provider and the user within a system or service context. For instance, certain parts of these SLAs may determine policies for the system resources that are provisioned to fulfill them. In order to make the SLAs operational, benchmarks need to be set and be based on logical entities that are “policy objects” which encapsulate their own rules.

On the overall, policy model can be useful in representing policy objects from different policies and correlating them. One such model is proposed in the IETF RFC 3060. The classes comprising the Policy Core Information Model (Moore et al, 2001), which are intended to serve as an extensible class hierarchy (through specialization) for defining policy objects. These will enable IE application (or service, comprising of several correlated applications) developers, IE infrastructure network administrators, and policy administrators to represent policies of different types. Policies can be derived for the various aspects of operations and management from the business objectives (Appleby et al, 2004). The service infrastructure (or part of it) provider needs to manage the policies (for that part) which determine how the environment is shared, costs are allocated, and so on. One important policy concerns with the diagnostic/repair servicing of the infrastructure itself; e.g. how to proceed when failure symptoms are observed. If a communications device fails between 8am and 9pm, call the system administrator; otherwise call the Help Desk.

The infrastructure provider, such as for network management services, defines the policies and parameters for its own services which will, in turn, define the parametric values of the end user service providers. Simply stated, if the mean time to repair a broken network is ‘ x ’, the end user services relying on the routers of that network cannot be repaired and brought back as live services in less than the time ‘ x ’. The IE service provider should define the key parameters for the services it offers so, at any instant, status and performance can be tracked and measured. Policies that share those parameters must be linked through a relationship in a policy model (such as the one mentioned earlier).

Current service quality management products do not support SLA compliance evaluations well because of their limited support of the adjudication processes for quality measures (Stern et al, 2002). In the UC arena, for IE, providers must proactively (a) maximize customer satisfaction

with competitive service level reports, (b) minimize the exposed business impact of service level violations, and (c) lower the cost-to-quality ratio of executing service level management processes.

6. DIAGNOSTIC STRATEGIES FOR THE IE INFRASTRUCTURE

What good is an IE if its designers don't think of its diagnosing aspect? Ubiquity and large scale imposes a requirement that service provisioning infrastructure should consider some sort of an autonomous way to diagnose itself to make it cost-effective. *Faults* and downtimes are unavoidable in communication systems. *Error* is a condition that is a consequence of a fault, that deviates the service from its established behavior. It is typically visible to the end users in the traditional environments, yet is expected to remain invisible in the IE. So, designs must take into account the high availability expected out of IE, and provide for the same. *Symptoms* are external manifestations of failures (which could include service performance problem with respect to an SLA). So, the IE must detect symptoms, trace the source of possible fault and diagnose the failure while continuing to provision the service with alternative resources from its reserve pool.

An IE may need to be managed by one or more applications, depending on its size, governance rules and other monitoring parameters. In large IEs spread over WANs, it is expected that performing fault localization process and maintaining the available knowledge base within a single management application may not be computationally feasible (Katzela et al, 1995; Wang et al, 1989; Yemini et al, 1996). Further, modern environments demand modeling of multi-layer, dynamically changing communication systems, and of representing uncertainty involved in dependencies between network objects (Steinder, 2001). This is because object relationships and policy executions between objects are dynamic and short lived as discussed by Duquenoy and Masurkar (2004) in the '*Maria – Road Warrior*' and other three scenarios of Aml.

One overall approach to take is to design automated fault localization technique that can improve performance of diagnosis not just in the lower protocol layers, but in all the the end-to-end layers, including application and service ones. In an IE functioning over a wide geographical area, service layer is realized by hop-to-hop services. So, in the '*Maria – Road Warrior*' scenario, getting a rental car which you may consider as an end-to-end service, may actually involve multiple hop-to-hop services. If a problem of a hop-to-hop service lies in the data link layer, considering that a large networked environment caters to a huge population, the number of

instantiations of such hop-to-hop services could be huge. Finding the affected hop-to-hop service to be relieved, replaced or fixed is complex. This underscores the importance of diagnosing multiple layers (Gopal, 2000). Steinder et al (2001) suggest a novel technique that allows to build a fault hypothesis in an iterative and incremental fashion and base fault propagation on Pearl's work on intelligent systems (1988).

It is expected that many new technologies will interplay in the IE; for example, biotechnology. In the earlier example, Maria wears a biosensor within the P-COM that sends signals to her physician who advises her to calm down right after her presentation. This implies that in order to maintain such devices with the infrastructure, understanding is needed of their interfaces into the information and rest of the infrastructure. In addition, behavior patterns of such devices such as in terms of performance profile, failure modes and diagnosability, to name a few critical areas.

7. CONCLUSION

Much more than traditional computing, UC facilitates strategic agility for providers of services for intelligent environments by making available computing services and business process components on a when-needed, as needed basis. As a business model, it continues to emerge in the industry, and there is potential to make it appealing to the end users of upcoming intelligent environments such as Aml.

There are many challenges, however, as mentioned in the paper. Particularly, in the areas of service pricing structures made uncertain by unpredictable market size and service adoption, dynamic policies needed for operations and management, and need for automated, intelligent fault management and diagnosis.

Diagnosing of the infrastructure to meet the high availability goals of the IEs is an aspect that needs to be thought about along with the building of the IE infrastructure itself. The size and complexity may decide, in turn, the fault and service management architecture. Certainly, there are many open questions in this area some of which are raised in the paper. These represent suggestions for further research.

In building computing models for providing services to IEs, different real life scenarios, and pros and cons of the different kinds of traditional vs. the UC models must be considered. Then it will become apparent why providers of intelligent services may have to continually evolve their businesses into an advanced type of the latter.

REFERENCES

- Appleby, K., Fakhouri, S., Fong, L., Goldsmith, G., Kalantar, M., Krishnakumar, S., Pazel, D., Pershing, J., Rochwerger, B., "Oceano – SLA Based Management of a Computing Utility, Proceedings of the 7th IFIP/IEEE International Symposium on Integrated Network Management, 2001.
- Appleby, K., Calo, S., Giles, J., Lee K., "Policy-based automated provisioning", IBM Systems Journal, Vol 43, No 1, 2004.
- Ducatel, K., M. Bogdanowicz, F.Scapolo, J. Leijten and J-C Burgelman. "Scenarios for Ambient Intelligence in 2010", *ISTAG (Information Society Technologies Advisory Group) Final Report, February 2001*, IPTS Seville, <http://cordis.lu/ist/istag.htm>
- Duquenoy, P., Masurkar, V., "Surrounded by Intelligence", in *Risks and Challenges of the Networked Society*, Proceedings of the 2nd IFIP Summer School, 2004 (Forthcoming).
- Gopal, R., "Layered model for supporting fault isolation and recovery", in Network Operations and Management Symposium [34], 2000.
- HP Utility Data Center, Technical White paper, Hewlett Packard Company, 2001.
- Katzela, I., Bouloutas, A., Calo, S., "Centralized vs distributed fault localization", in IM'95 [20], pp. 250-263.
- Moore, B., Ellesson, J., Strassner, J., "Policy Core Information Model, Version 1 Specification, Internet Engineering Task Force, IETF RFC 3060, <ftp://ftp.rfc-editor.org/in-notes/rfc3060.txt>, 2001.
- N1 Grid Service Provisioning System 4.1, Sun Microsystems, Inc., http://www.sun.com/software/products/service_provisioning/index.html
- Open Grid Services Infrastructure Working Group (OGSI-WG), Global Grid Forum, <http://www.ggf.org/ogsi-wg>.
- Paleologo, G., "Price-at-Risk: A methodology for pricing utility computing services", IBM Systems Journal, Vol 43, No 1, 2004.
- Pearl, J., "Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference", Morgan-Kaufmann Publishers, 1988.
- Rappa, M., "The utility business model and the future of computing services", IBM Systems Journal, Vol 43, No 1, 2004.
- Steinder, M., Sethi, A., "The present and future of event correlation: A need for end-to-end service fault localization", Proceedings of the 5th World Conference on Systemics, Cybernetics and Informatics, 2001.
- Stern, R., Erickson-Harris, L., Onge, D., "SLM Solutions: A Buyer's Guide", Enterprise Management Associates, Boulder, CO, 2002.
- Sun Microsystems, Inc., "The I-Fabric Architecture", 2002.
- ThinkDynamics: On Demand Provisioning, IBM Corporation, 2003.
- Wang, Z., "Model of network faults", in IM'89[17], pp. 345-352.
- Yemini, S., Kilger, S., Mozes, E., Yemini, Y., Ohsie, D., "High speed and robust event correlation", IEEE Communications Magazine, 34(5):82-90, 1996.