

## **The Anatomy of Large Scale Systems Revisited**

**Joel Moses**<sup>1</sup>

*MIT, Cambridge, Mass., 02139*

Copyright © 2009 by J. Moses. Published and used by MIT ESD and CESUN with permission.

### **Abstract**

*In papers at the prior ESD Symposia we discussed system properties, often called ilities, as well as the structure of large scale systems. Issues related to ilities have been relatively well understood, but issues related to structure and organization were not. Hence this paper addresses this issue again. The main architectural concept in the earlier papers was layered hierarchies. We will discuss this concept in the context of applications, such as health care and higher education.*

### **1. Introduction**

I wrote a paper for the 2002 ESD Internal Symposium entitled “The Anatomy of Large Scale Systems” [1]. I also wrote a paper for the first ESD Symposium in 2004 entitled “Foundational Issues in Engineering Systems: A Framing Paper” [2]. The latter paper discussed among other issues the notion of ilities and related system properties and was a lead paper in the ESD Monograph. Other Monograph papers discussed architecture, enterprises, uncertainty, safety and sustainability. In retrospect, the importance of the ilities and related properties of large scale systems was clearly understood by the Engineering Systems Division’s faculty. Unfortunately, the key issue in the Anatomy paper and a related paper entitled “Three Design Methodologies” [3], which was about the structure of certain large scale systems was not well understood. Hence this paper addresses that issue again.

What I have learned over the years is that most engineers have a different understanding of architecture than computer scientists, in particular computer scientists with a background in programming languages as well as practice with modern large scale software systems. The key architectural concept in the Anatomy paper is that of layering. I will describe it again, try to explain why mechanically oriented as well as many other engineers do not appear to understand its importance, and consider how it might help explain new architectures for major industries, such as health care.

---

<sup>1</sup> Institute Professor, ESD and EECS, Room 32-249

I became a graduate student in Artificial Intelligence in 1963. The key approach to AI in that era was Heuristic Search. This approach assumed that each problem could be converted to a top-down tree of subproblems. That is, you take the original problem, for example a theorem in logic, and break it into parts. Choose a subproblem (via heuristics) to break down further until the subproblems become so simple that one can readily solve them and back up the solutions in the tree of subproblems until one has arrived at the solution of the original problem. This worked reasonably well in proving simple theorems in areas of mathematical logic in the 1950's [4]. The approach eventually led to a world-class chess machine in the 1990's. In many other areas, such as understanding of sentences in a natural language, the approach did not succeed very well.

I began to be concerned over the Heuristic Search approach by 1965. Did people really believe that one could, for example, develop the general theory of relativity without having specialized knowledge of physics and mathematics? In my thesis in 1967 [5], I introduced the concept of Systems with Expertise, later known as the Knowledge Based System approach to AI. I argued that a top-down approach would generate immense tree structures for sufficiently difficult problems, tree structures that could not be searched in the lifetime of our universe. On the other hand, humans rely on knowledge developed over many years by many people, and one should simply use such knowledge as best one can. Some AI researchers did not like the idea of introducing specialized knowledge then and even now, but many others felt that this was the way to go further in AI, and the approach became popular for a while. There was another point that I did not emphasize in my thesis, partly because it was obvious to me – the knowledge base had to be *structured* in order to be useful in problem solving.

In the late 1970's I became a middle manager at MIT, namely a department head. I decided I should read some of the literature on human organizations. I found out that one of the most referenced books was by Herb Simon, a key figure in AI in the 1950's. Herb suggested thinking of organizations as tree structures [6]. I became quite frustrated and could not understand why this top-down approach was being used in so many areas, when it was clear to me that while it had many advantages it also had serious weaknesses, in particular in having low flexibility relative to the size of the system.

In the early 1980's a growing concern in the US was on manufacturing. It was increasingly clear that US companies were losing out to Japanese and German ones in a variety of industries, but especially in automobile production. I started reading the literature on Japanese manufacturing and found a book, Theory Z by Ouchi [7], that described an approach to Japanese organization that I recognized as being close to the one I had advocated in AI. I have called the organizational structure that I had been advocating layered hierarchies as opposed to tree structured hierarchies. In a hierarchy there is an explicit or implicit rank ordering. Your boss ranks higher than you. A problem ranks higher than any of its subproblems. A key difference between tree structured and layered organizations is how nodes (e.g., subproblems, people) of the same rank or level are related to each other. In a tree structure there may not be much interaction between sibling nodes. In a layered organization sibling nodes interact with each other and likely will cooperate. Cooperation versus competition is a critical difference between these two organizational structures, and reflects differences in the ideology of the people that create the system architectures based on them. The US arguably has greater

emphasis on competition than any other major nation. This accounts, we claim, for the prevalence of tree structured organizations in the US.



Figure 1 Layered Structure with Three Layers

Elements in one layer can connect to one or more elements in a higher layer. Often members of one layer interact closely with other members of that layer.

A question that has concerned me for the past few years is why many engineers, mostly engineers other than EEs and CSs, do not use layered structures in their designs. Systems Engineering and Software Engineering both rely on tree structured organizational frameworks. Yet many large software systems and even many small ones do use layered approaches, which have the property of keeping overall system complexity in check. I do not think that competition versus cooperation in the design teams is sufficient to explain this phenomenon. My current answer is that engineers are highly influenced by the devices and systems that they can design. If the systems lend themselves to abstraction as well as horizontal or lateral (as opposed to just vertical) interconnections, then layering makes sense. Unfortunately most engineering devices and systems do not have both of these properties. Software and digital hardware have these properties, and thus layering is frequently found in such systems. Human organizations also can have these properties, and thus layering can be found in both human organizations and industries.

One may wonder where these two organizational structures first appeared in written form. It is not a joke, but they seem to first occur in the Bible. My namesake, Moses, had an important part in this story. His father-in-law, Jethro, went to meet Moses at the foot on Mount Sinai, and brought back to him his wife, Jethro's daughter, and his children. Jethro saw that Moses was spending all his time rendering judgments in disputes among the Israelites. He told Moses that this was not good. Moses should appoint a judge for every thousand, every hundred, every fifty and every ten, and this way he would only have to render judgment on the most difficult cases. Moses said that he would make these appointments.

What Jethro was proposing was a tree structured organization with Moses at the top node of the tree. Disputes went from a bottom node up the tree if they were sufficiently difficult cases. The most difficult cases of all would be judged by Moses himself. Whereas our examples earlier of tree structured organizations were top-down, this structure is bottom-up. Nevertheless, both are tree structured organizations with each node other than the top one having exactly one parent node. In retrospect this is an obvious solution to a situation where a person is overwhelmed by too many inputs. Why did Moses not see it before it was suggested by Jethro? My answer is that Moses did not think in the way Jethro thought. Moses' approach to organization can be seen in the book Numbers.

In Numbers Moses describes the religious organizational structure of the Israelites. He appoints his brother, Aaron, as High Priest, the top of the religious hierarchy. Aaron's sons become the top level of the hierarchy, and their male descendents all become priests. The rest of the tribe of Levi and their descendents become the middle layer. The remaining Hebrew tribes become the lower level of the hierarchy. Thus we have a hierarchy with three layers. Maimonides later explains how the priests and Levis performed their functions [8]. The priests operated in teams, accepting sacrifices from Israelites in the Temple. Israelites were helped by a nearby priest. They did not have an assigned priest. The Levis also worked in teams helping priests in their functions or playing musical instruments and singing. Levis also did not have pre-assigned priests. This shows the cooperative aspects of a layered hierarchy.

Tree structured organizations lend themselves to competition between sibling nodes. Sloan's organizing principles for GM in the 1920's had the various auto divisions (e.g., Chevrolet, Buick) compete for resources within the company, but not on price. As a result they did not compete for customers. An obvious question, given the two types of hierarchy, is which one is best. The classic answer is – it depends. In some cases competition within the organization is very important, and then tree structured organizations might be best. In other cases cooperation within the organization is very important, and layered hierarchies might be best. There are other issues that affect the choice. If the rate of change within the organization and in its environment is low, tree structured organization might be better than layered ones because there usually is a nontrivial overhead associated with going from one layer to another. If the rate of change is very high, neither architecture is very good – a networked organization is usually better in such circumstances. We claim that in intermediate values to the rate of change layered organizations might be better. We do not believe that there is an organizational structure that is ideal under all circumstances. As we shall see, each of the major architectural approaches is good in some important cases, and each is not best in certain other cases.

### *1. A Layered Systems and the Magic Number 3*

Many of the examples in the Anatomy and Methodologies papers were of technical layered systems, such as large software systems or mathematical ones. For example, the microprocessor in my PC, the database system that uses it, and the database application that uses that system form a layered system with three layers. Here we emphasize human organizations and especially industries that are or can be layered.

Layered human organizations tend to have three layers. Hence we call 3 the magic number of layers. Some layered organizations have 5, 7 or 9 layers. This often occurs because a single layer was decomposed into three layers one or more times, thus adding two to the total number of layers with each decomposition. Due to human processing limitations most large scale organizations are actually hybrid ones, using tree structures globally and layered or team structures locally. Figure 2 presents an example of such a hybrid organization.

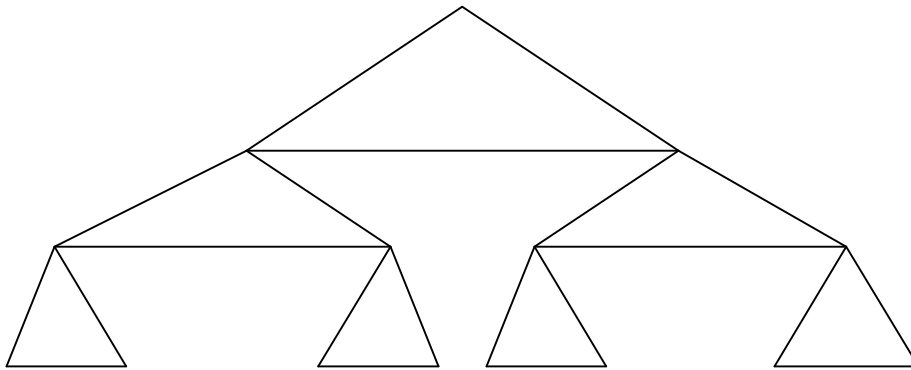


Figure 2: A Hybrid Tree and Layered Organization

The religious hierarchy in the Bible had three layers (priests, Levites and Israelites). Plato created in *The Republic* [1] a Just Society that had three layers as well. The guardians were in the top layer (also called the gold layer). Their leader was the Philosopher-King. His structural position was comparable to that of Aaron, the High Priest. Below the guardian layer was a silver layer of people, and below them was a bronze layer. Plato and Moses had enough in common that the second century philosopher Numenius said that Plato was simply Moses speaking in Greek [11]. A key difference between the Just Society and the religious hierarchy in the Bible was that in the Bible men were born into a given layer and could not advance from it. Plato relied on education and continual testing of both men and women to determine the layer to which one belonged. Plato believed in cooperation so much that he created a communist state for the guardians. They shared everything, including wives. Most people would agree that Plato went too far in this characteristic of the guardians.

The Catholic Church is the oldest existing institution in the West. Its structure and its early teachings are based on Neo-Platonism. The Pope is structurally the Philosopher-King. Cardinals or Archbishops are the top layer of the hierarchy at the present time. Cardinals work together to elect a new Pope. Bishops compose the middle layer and priests the lowest layer. The structure of the Church had a significant impact on the structure of universities. University rectors or provosts are the philosopher-kings. Full professors are the top layer and are the archbishops or cardinals of the university. Associate professors are the bishops and

assistant professors are the priests. While the importance of this layered structure of universities has much diminished as a result of the introduction of research-oriented departments in the 19<sup>th</sup> century and the growth of university administration, it is not completely gone. The tenure process whereby professors move from one layer to the next higher one shows remnants of this organizational structure.

Tree structured organizations are most common in the US, but layered organizations abound. Most large partnerships are layered. Large law firms are usually composed of senior partners, junior partners and associates. Making partner in such a firm is akin to obtaining tenure in a university.

Another structure influenced by the Church is the structure of masters, journeymen and apprentices in Middle Age guilds. Apprentices worked for a number of years for a given master. As they became better at goldsmithing, say, they graduated to become journeymen. Journeymen moved, that is journeyed, from one town to another working for one master after another. Experienced journeymen could present a master piece to prove their ability, and then could join the ranks of masters in a guild. Germany followed this approach in some fields as late as the 20<sup>th</sup> century. My father was apprenticed to a competitor of his father in Germany in the 1920's in order to learn the business of buying and selling scrap metal.

Note that each example above has exactly three layers. This is not an accident. A large system with exactly one layer is not hierarchical. It is best described as a network of component parts. Hierarchies can improve the effectiveness of the system by allowing people to make relatively rapid decisions in complex cases. Hierarchies are also useful in cases where the subordinates do not agree as to the best course of action. A system with exactly two layers is hierarchical, but does not enjoy the power of a system with three layers. Simple multi-celled biological species can have one neuron, but do not have the behavioral complexity of species with three or more layers of neurons. Large layered software systems tend to have lower complexity and greater flexibility than other hierarchical structures that have similar behavior.

Ouchi's Theory Z book made me realize that large Japanese firms can be organized in layers. Outwardly these firms appear to have a tree structured organization. Yet there is an underlying layered hierarchy that is to a first order based on age. The number of such layers is not limited to three, but can be five or seven, depending on the size of the organization. Middle managers in such organizations work to develop trust among their staff and those of other middle managers at the same layer. This is done so that members of new teams will be able to work together quickly and effectively.

## **2. Organization of the US Health Care System**

Health Care is an excellent example of a domain that can be organized in a layered fashion. It is an area where the government plays an important role. In particular, governments around the world pay a significant part of the health care costs in their nation. In the US, Medicare and Medicaid cost a significant fraction of the overall US health care bill of about \$2.4 trillion. Thus the cost of the health care system is an important national issue. Most

governments also wish to have good health care for their nation, and individuals wish to have the highest quality health care for themselves and their families. Thus, there is usually a struggle between the nation's needs and ability to pay for health care and individuals' wants in this domain. This struggle is sometimes resolved by having both a national system of payments and a private one that supplements it. In layered terms the struggle is between individual desires to use services at a high layer and government's desire to rely on health care services at the most appropriate layer, which may be lower than the individuals feel that they need or deserve.

The current US architecture of health care can be best described as a hierarchy having two layers. The bottom layer is that of primary care practices. The top layer is that of hospitals and specialists. There are other health care organizations, such as nursing homes, but we view these as separate from the main parts of the system. In many European countries the overall health care system is a hierarchy with three layers. This includes a bottom layer which is composed of community clinics, largely run by nurses. Individuals expect to go to the local clinic for many of their symptoms, and expect to be referred to a physician or the emergency department of a local hospital if the situation so warrants. Nurses in these community clinics are empowered to perform a sizable percentage of cases that are presented to the health care system. Some limitations on the ability of nurses to treat patients is clearly useful and needed, but careful empowerment of them can improve the overall performance of the health care system.

There are advantages and disadvantages to having a first layer of community clinics. Doctor's offices are usually closed at night, whereas the community clinics are usually open at late hours. Doctors have increasingly tended to avoid visiting patients in their home. Nurses could do that and check on elderly patients and those who are chronically ill. Many of the presenting symptoms to a primary care physician tend to be fairly straight-forward and could be safely handled by a nurse practitioner. The current US situation makes the job of a primary care physician somewhat boring, although the payments for such visits may be necessary for the income of the primary care practice. Primary care physicians are also paid less than many specialists, and this adds to their frustration with the job. The boring nature of the job and the relatively low compensation are among the reasons it is difficult to attract young doctors to careers in primary care. From a national perspective a heavy reliance on physicians as the first layer in health care is overly expensive.

The disadvantages of having nurse practitioner-based clinics as the first layer of the health care system include concern over patient safety. This can be alleviated by restricting the cases that are handled in such clinics, as well as having physician oversight of the clinics. Increasing reliance on nurse practitioners will require an increase in the number of nurses and will place strain on the nursing education system. On the other hand, the number of primary care physicians is decreasing for the reasons noted above, and a change such as we discuss here is consistent with that reduction.

The layer of hospitals and specialists can itself be broken into three layers, thus making five layers overall for the health care system. Major teaching hospitals are often called tertiary hospitals, implying three levels of hospitals. The first level or layer is composed of

community hospitals which can handle many relatively simple cases requiring generalists, specialists and surgery. More complex cases would be sent to a regional hospital, if there is one, or to a tertiary hospital. Tertiary hospitals handle cases that have been difficult to diagnose. These hospitals also have sophisticated intensive care units, several sub-specialists and relatively complex machines.

In some parts of the US there are specialized hospitals that handle, for example, heart disease, cancer, diabetes or children's diseases. The advantage of such specialty hospitals comes from the experience gained by teams of doctors, nurses and other staff in seeing many similar cases. Such teams can improve their ability to treat such cases over time. With higher quality outcomes, there should be fewer complications and the overall cost of each procedure should be markedly reduced. The discussion of specialized hospitals is influenced by recent books of three Harvard Business School faculty [12, 13, 14].

We propose a three layer structure for tertiary hospitals that builds on an intermediate layer of specialty subhospitals. The lowest layer of these tertiary hospitals would have an ER and some general medical services, but fewer such services than at most present tertiary hospitals which tend to duplicate services available at other hospitals. The middle layer would have several specialty subhospitals. These would share some medical services (e.g., anesthesiology) as well as other services (e.g., HR, IT). The top layer of these tertiary hospitals would handle the difficult to diagnose cases or very complex procedures. Master diagnosticians, such as TV's House minus his personality quirks, would practice in this layer. Teaching and research should also play a key role in this layer of a tertiary hospital.

A key advantage of specialized subhospitals is that, as we noted, with much practice teams of doctors, nurses and other staff members can continually improve their ability to treat a class of patients with high quality at relatively low cost. This is consistent with the Toyota approach to manufacturing where manufacturing teams get better over time through a process of continuous improvement. Physicians in the specialized subhospitals may need to play multiple roles. They are, of course, members of specialized teams in a subhospital. They are also specialists who may need to consult on cases occurring in the emergency department or the tertiary part of the hospital. Some specialists, such as pathologists or radiologists, may need to consult in several subhospitals as well as the other two layers of a tertiary hospital. A given tertiary hospital need not have all possible specialized subhospitals. Some specialized subhospitals simply do not have the volume that justifies their existence in many tertiary hospitals. The competition between subhospitals in the same specialty should lead to some having large volume based on cost, quality and general reputation, thus using competition to drive out other subhospitals in the same specialty.

The main goal of the top part of the hospital is to help the patient presenting symptoms that are difficult to diagnose or correct. Here we would expect that cost and quality can only be secondary goals. This is the place where master diagnosticians practice and rely on other physicians and staff on an if-needed basis.

Having a clear separation of the three layers of a tertiary hospital has the advantage of reducing the complexity of the overall hospital. Such a reduction of complexity should reduce

the overhead and the overall cost of running the hospital. Christensen [12] points out that there ought to be a clear difference in how the specialized subhospitals are paid from the way the top layer is paid. The specialized subhospitals ought to be paid a fixed price for each procedure. If there are complications associated with the procedure in a given patient, the specialized subhospital will have to take care of it without additional payments. The top part of the hospital ought to be paid largely by the time it takes to treat a patient and all the services required. This is close to the current payment method, and results from the difficulty of diagnosing and treating these patients.

### **3. Other Three-Layered Organizations and Industries**

The third layer in our proposed tertiary hospital deals with complex cases that require master diagnosticians. The middle layer in such hospitals involves specialized hospitals with teams that improve with experience. Such teams not only increase the quality of the outcomes but can reduce the overall costs. The lowest layer in the health care system which involves nurse practitioners should deal with situations that are so well understood that the diagnosis and treatment can be made at even lower cost with near certainty of outcome.

Many other systems have a similar hierarchy. Consider women's fashion. At the high end the clothes from a salon in Paris are expensive, say \$25K per dress. One pays for the creativity and reputation of the master designer. Often the materials in these outfits are quite expensive. The clothes will have an excellent fit that comes, in part, from several visits to the salon for each dress. One does not become a master designer overnight, although nowadays the guild system is no longer in operation. These dresses are largely hand-made by a number of seamstresses, some of whom may be acting as apprentices for the design studio. Of course, the designer may not have had his or her best season, and thus the dresses may not work well that year.

At the middle level of the fashion industry one can get dresses similar to recent showings in the salons. Such dresses might cost about \$3k and would usually be based on a single fitting. Presumably the knock-off dresses are ones that have been viewed by several eyes as being among the best for that season. The lower level would have dresses that are significantly cheaper and each fitting change would involve additional expense for the buyer. These dresses might not be related to what is being shown in the salons that year. Much automation and low cost labor is involved in the production of these dresses.

Consider the automobile industry. At the high end one can have special cars that cost upwards of \$100k. These are largely designed by a master auto designer and may be largely hand-made. These cars often have high performance and have expensive materials but may spend a fair amount of time in the shop. Low-level cars will usually be small and have relatively few features. Middle-level cars will usually be larger than the low-level ones and will have many more features. These cars will also compete on quality as well as cost. The Toyota Production System can result in high quality at reasonable cost for such cars [15]. Cars in this level may have fewer defects than some of the high level ones.

In the private sector one usually relies on price to determine the layer in which one makes purchases. Indeed, one might not want to purchase high level dresses or cars because purchasing them might take a long time, and one has to handle the product with great care. But to a first order one will purchase the product at the highest level one could afford, at least some of the time.

The public sector plays an important role in many fields, for example in the military, education as well as health care. In these fields individuals want to have the highest level of service, service that will largely be paid for by the government. The government, however, must pay attention to overall costs. Hence it will tend to emphasize obtaining services at the lowest applicable levels.

A layered approach to education in the US shows up at the college level, in particular in California. Community colleges are the lowest layer of the system. The California State institutions form the middle layer, and the University of California campuses form the upper layer. The major private universities, such as Stanford and Cal Tech, are at the upper layer as well. In the upper layer professors show their mastery of the educational material by doing research in that area or one closely related to it. The hiring process of these faculty members also assures a level of mastery. As noted earlier, university faculties have been themselves organized in three layers in past centuries.

The admissions process determines to a large degree into which layer a student enters. The cost to the state on a per student basis usually becomes higher as one moves up the layers. Parents can spend additional moneys to have their children go to private institutions, which presumably offer a premium level of education or at least provide an increased cachet at any given layer. Ideally, education at the lowest layer, namely community colleges, should be of high quality although possibly not at the breadth of liberal arts schools nor offered by research faculty members. It is not clear that this level of quality has been or can be achieved at this time.

Education at the K-12 grades is influenced by national ideology to a considerable degree. Thus the layering we note at the college level can be absent. There is less emphasis on private education at any level in continental Europe than in the US, for example.

The military is a tree structured organization. Cooperation between services, such as the Air Force and Army, can, however, lead to structures that are akin to layering that rely on lateral alignment. The DoD has wanted the Air Force and Army to cooperate more closely for decades. The army wants the air force to provide support to its troops by bombing enemy targets ahead of the ground forces. The targets would be chosen to be close to the ground forces, but not so close as to potentially hit our own men. The air force hierarchy often still believes that they could win the battles all by themselves, and would prefer to shoot at targets far in advance of the army and largely independent of the army's immediate needs.

Close cooperation between the army and air force could reduce the overall need for ground troops, and may result in superior military results. Such cooperation will be needed in several levels of the military hierarchy, at strategic, operational and tactical levels. Analysis of recent

wars engaged by the US military indicates that such cooperation at multiple levels of the military hierarchy was finally achieved in the early phase of the war in Iraq in 2003 [16].

## References

- [1] J. Moses, The Anatomy of Large Scale Systems, <http://esd.mit.edu/WPS/ESD-WP-2002-05.pdf>, pp.49-51, 2002
- [2] J. Moses, Foundational Issues in Engineering Systems: A Framing Paper, <http://esd.mit.edu/symposium/pdfs/monograph/framing.pdf> , 2004
- [3] J. Moses, Three Design Methodologies, <http://esd.mit.edu/symposium/pdfs/papers/moses.pdf> , 2004
- [4] A. Newell, H.A. Simon, The Logic Theory Machine, RAND Corp., Santa Monica, Calif., 1956
- [5] J. Moses, *Symbolic Integration*, MAC-TR-47, Project MAC, MIT, 1967
- [6] J.G. March, H.A. Simon, *Organizations*, J. Wiley & Sons, 1958
- [7] W. Ouchi, *Theory Z*, Addison-Wesley, 1981
- [8] Maimonides, *Mishneh Torah*, volume 8
- [9] A. P. Sloan, *My Years with General Motors*, Doubleday, 1964
- [10] Plato, *The Republic*, transl. D. Lee, Penguin Books, 1951
- [11] Clement of Alexandria, *Stromata*, i. 342
- [12] C. Christensen et al, *The Innovator's Prescription*, McGraw-Hill, 2009
- [13] R. Herzlinger, *Who Killed Health Care?*, McGraw-Hill, 2007
- [14] M. E. Porter, E.O. Teisberg, *Redefining Health Care*, Harvard Business School Pub., 2006
- [15] J.P. Womack, D.T. Jones, D. Roos, *The Machine that Changed the World*, HarperCollins, 1991
- [16] J. Dickmann, *Operational Flexibility in Complex Enterprises*, Doctoral dissertation, ESD, MIT, June 2009