Executive summary

ICT SHOK Future Internet programme forms a vital, integral part of the global research efforts taking place in Europe, US, China, and elsewhere. It focuses on enhancing the networks’ ability to reduce transaction costs and promoting co-operation, thereby creating new wealth and supporting democracy. Hence, the long term goal is to make information networking a reality, with the shorter term goals leading to it.

The two major undergoing changes in the Internet is the move from the end-to-end principle towards a trust-to-trust principle, at the same time moving from interconnecting nodes towards interconnecting information. While doing this, scalability and autonomous resilience remain the main challenges to address.

There are six main present problems: unwanted traffic, choking of the routing system, mobility and multihoming, compensation and congestion, privacy and attribution, and trust and reputation. The major future challenges to properly address, in addition to the already mentioned overall question of information networking appear to be energy consumption, usability and usage patterns, and increasing core capacities.

At this phase of writing this document, we dare not to postulate a concrete vision or mission. We expect them to emerge at a later phase of the writing. However, there is some initial drafting later on.

To reach the drafted vision, we need at least four major new components added to the network architecture: mechanisms for explicit representation of reputation and trust, strong identities with a suitable balance of privacy and attribution, new compensation mechanisms that disentangle the various functions of money, and new structured market places.

Introduction

The Future Internet — just like the Internet today — will be a global network. Consequently, doing research on the area will necessarily be a collaborative effort, focusing heavily in drawing ideas from the rest of the world and pushing aggressively our own views and innovations abroad. Consequently, the ICT SHOK Future Internet programme shall not be viewed as a standalone research effort but as a vital, integral part of the wider research efforts taking place elsewhere in Europe, as well an in the US, China, and the rest of the world. In practical terms, our existing, well establishes connections to the NSF NeTS FIND and GENI programmes, the related EU FP7 projects, and contacts with Tshinghua university and the Chinese next generation internet programme must be tightly integrated within the ICT SHOK umbrella, and strengthened with strong, personal-level contacts with other major Future Internet research programmes around the world. Focusing solely on what happens in Finland would be a dire mistake.

As the name says, the Internet is a network of networks. Its main benefit is in social and economic networking, i.e., in reducing the transaction costs when communicating with other people and businesses. From that point of view, in the Future Internet programme the main point to investigate is the networks’ ability to deliver useful services to the users. The work shall be tuned towards systems where “the user is happy and satisfied”, i.e., where the user considers getting value for his money. Taking a more economic angle, we can define these current and future services as things that provide more utility to the users than is required for their production by their providers, thereby increasing the total amount of wealth in the world.

The long term goal should be a network where “all of the magic” happens automatically, without human action needed. From that point of view, we can say that the vision is to create the networking version of Adam Smith’s “invisible hand” that guides the markets towards increasing wealth, combined with an open, democratic, socially fair approach that will force the created wealth to be distributed more evenly than what a raw, pure capitalism would induce.

As discussed in more length later, the long term goal should be to make information networking a reality. The short and medium term goals should be some easily identifiable goals which lead to the long term goal. For producing immediately beneficial, short term results, there are a few research groups amongst the planned participants who have worked on peer-to-peer content distribution and other peer-to-peer based system. This, combined with the ground breaking word on the current Internet architecture could be used as short term goals; the expected results there are pretty much known and mainly need qualitative, measuring research rather than ground-braking innovations.

The rest of this document is organised as follows. First, we outline the main background for the currently perceived need for change in the Internet. After that, we discuss a number of current major problems in more detail. In the next section, we outline a few anticipated major future challenges. Finally, in the last section of this version, we move towards a joint vision. However, it must be understood that this version of the document is very much work in progress.
Background

The original Internet was designed for connecting computing devices into a global web of computers. Ever since its dawn, the advancements of the computing technology and its applications have guided the evolution of the Internet. The current computing trends with a major impact to the Future Internet include the following: mobile computing is transforming mobile phones into multimedia computers that need high bandwidth access to content, Web 2.0 applications build on the assumption of always available connectivity to network embedded storage and server applications, and the emerging sensor applications are broadening the scope of the Internet into physical world. The sphere of the Internet is continuing to increase. Voice and data (Telco and IT) convergence is setting the capacity, performance and usability requirements onto a new level as the Internet is becoming a central part of infrastructure of all societies. New business models are emerging that build on the access independency, global reach, huge existing content, application and developer base of the Internet. This business models will introduce their own set of availability and accountability expectations that we not in the scope of the current design of the Internet.

From end-to-end to trust-to-trust

The fundamental design principle of today’s Internet is formulated in the original work on the end-to-end principle. The principle seemingly stipulates the execution of functionality in endpoints of communication, following secondary principles like minimality, generality, simplicity, and openness.

When looking at many deployments of communication services in today’s Internet, it becomes apparent that placement of the execution of functionality is crucial when designing solutions to be deployed. This, in effect, has led to the appearance of all kinds of middle boxes, such as firewalls, NATs, but also services that are largely implemented in trusted 3rd parties rather than the endpoints of the actual service (from a user’s perspective).

The realisation that the place of executing functionality properly and in a trustworthy manner is crucial when designing an overall solution to the given communication problem, has led to a line of thought brought forward by David Clark in his attempt to clarify this point in the original formulation of the E2E principle. As a result, a trust-to-trust (T2T) principle has been created to capture the essence of this, as follows:
“The function in question can completely and correctly be implemented only with the knowledge and help of the application standing at points where it can be trusted to perform its job properly.”

For instance, looking at a firewall, in the light of T2T, as a component that implements traversal reception in a larger enterprise network in a trustworthy manner (from the perspective of the enterprise) makes a firewall part of the entire end-to-end solution rather than pinpointing it as a hack that was inserted after the fact.

However, there is another, more subtle, shift that occurs when moving from E2E to T2T. The original end-to-end principle has, not only by its name, a strong focus on the notion of endpoints. There is an underlying, usually implicit assumption that the recipient is willing to receive whatever the sender is sending, or else the sender would not send in the first place or will cease to send very quickly. Obviously, this assumption is no longer true.

As stated before, the careful clarification of E2E — in its form of trust-to-trust — seemingly shifts the focus away from (network) endpoints to (trustworthy) points of execution. This opens up a discussion whether or not a network-endpoint-oriented paradigm, like IP, is well suited under the refined T2T principle. Also, as argued, for instance by Van Jacobson, many communication scenarios aren’t even focused on endpoint communication at all rather than the retrieval of information from (trustworthy) sources that can provide the information. This observation, together with the T2T clarification, seems to call even stronger for revisiting the design of the network endpoint oriented IP solution.

Against this background, one cannot overemphasise the importance of new, controlled openness. Only by moving towards a network that is simultaneously open and genuinely trustworthy one can efficiently support new application and services. In order to reach such a situation, it is necessary to understand the economic and other tensions between providers and consumers in one hand and between different types of providers in the other hand. Such tensions must be revealed and opened up, preferably in the form of creating new open and level market places, so that the tensions can be internalised in the form of market-based prices or other compensation mechanisms.

**Scalability — the fundamental issue**

The Internet is one of the most complex machines humanity has ever built. In the future, when moving towards more ubiquity and shifting attention from computing platforms to information objects, the level of complexity is likely to grow by a number of magnitudes. Hence, the network must be built from the beginning with scalability in mind. Everything must scale; therefore, we do not explicitly mention scalability below but expect that each and every aspect of the network must necessarily scale. To give some ballpark figures to get an understanding of the scale of the complexity, we assume in the order of $10^{10}$ autonomous domains and order of $10^{15}$ information objects.

While apparently easy, scalability is, in practise, XXX.

**Some lesser issues**

Other crucial overall challenges are mostly related to the changing nature of networking. Instead of being a mostly-XXX

- **Zero-configuration; Personal zone / space**
- **Moore’s law becoming to the end.** No more faster CPU cores, moving to multi-cores.

### Current problems

**Unwanted traffic**

The various forms of unwanted traffic, including spam, distributed denial of service, and phishing, is arguably the biggest problem in the current Internet. The root reasons to unwanted traffic seems to be best characterised with economics. We can characterise the current Internet as a global, distributed system where the main cost of communication is paid by the recipient. This is a direct (though certainly unintentional) consequence of the network architecture. By explicitly and directly naming all the potential recipients, we create a system where the senders can easily express their desire to send data to any recipient in the network. Given that under the typical contracts the marginal cost of sending additional packets is very close to zero (up to some capacity limit), there are few or no incentives for refraining from sending unwanted traffic; sending some packets just for fun costs so little that it doesn’t matter. At the same time, even a mar-
A related but slightly different phenomenon is the basic communication efficiency, i.e., the fraction of signalling and other overhead vs. payload. XXX expand.

The current practice to fight unwanted traffic is to add more middle boxes. However, they cause a number of different reachability problems, creating a significant challenge in today’s network design. There would be need to develop some form of controlled reachability; i.e., how to prevent the unwanted traffic from consuming network resources, especially over wireless link, but at the same time allow flexible use of different servers, services and protocols also behind the wireless links.

- Near term: protocols to interact directly with NATs and firewalls; unwanted traffic on other layers, such as spam mail or unwanted IM spam messages
- Medium term: communication frameworks for preventing unwanted traffic arriving from upstream; end-hosts to signal what services they provide and what traffic they are interested in.
- Long term: economic solutions; completely new architectures.

**Choking of the routing system**

The second big problem can be described as relative choking of the routing system. The current Internet routing system relies solely on the Border Gateway Protocol (BGP); a protocol that has received some facelifts but internally has remained the same for the last decade or so. At the same time, the business environment where the Internet Service Providers (ISPs) compete has become immensely more complex and competitive. From the operational point of view, the main technical term characterising the current routing complications is traffic engineering. A basic fact is that the current routing system, i.e., primarily BGP, does not offer any good facilities for it — almost all of the various ways that the ISPs attempt to perform traffic engineering can be considered as some sorts of protocol violations or hacks relying on obscure side effects of the routing mechanisms.

A potentially even bigger problem appears to be routing table converge. For example, there are indications that the routing system may fluctuate days or even weeks after major events affecting the links, such as a recent undersea earthquake near Taiwan that cut a handful of undersea communication cables.

At the same time the non allocated IPv4 address space is projected to exhaust in some two-to-three years, creating urgency for migrating to IPv6. The transition period will stress more the routing infrastructure as two different address spaces need to operational. Consequently, the Internet research community has started develop new approaches to meet the scalability and traffic engineering needs of the core Internet.

Altogether, these issues should be taken as early warnings, indicating that our current routing system may be near its inherent capacity limits. With further growth, the ISPs may no longer be able to perform effective traffic engineering, probably leading to some market consolidation and loss of competition, and the routing system may start to experience global-scale instabilities making large fractions of the Internet unavailable for excessive time periods.

**Locality** - obviously the use of the backbone costs money. So if operators can make most communication local even when roaming, the better. This would be infrastructure provided route optimisation with services level intelligence.

Specific research items in this area include the following:

- Near term: IPv6 migration; Renews role of NATs; BGP housekeeping.
- Medium term: Extended link layers; cross-layer interactions; pressure from mobility and multi-homing; pressure from overlay networks.
- Long term: Compact routing; Hierarchical or recursive routing; Understanding the consequences of power law or long tail topologies.

**Mobility and multi-homing**

The current Internet architecture was not developed with mobility in mind. Today, there is clear need for combined mobility and multi-homing support at different levels of granularities ranging from users and applications to complex subnetworks. Effective mobility support requires a level of indirection, something that the current Internet architecture is gravely missing at most potential mobility points. Adding suitable indirection points is likely to require architectural changes.

Already today, a typical wireless end-host can use several network access technologies to get connected to the network. In addition, rather than having just single “Internet”, the world consists of a number of isolated network islands in addition to the publicly open Internet. A wireless terminal may see a number of available network access points using different network technologies, and without making an initial attachment by which it gets additional information of the network characteristics a wireless host has very little information about the identity and characteristics of the networks behind the access point. A further
complexity is that in addition to the physical access interfaces, different types of virtual interfaces or VPNs can be used to access a specific network domain.

Furthermore, many of the current services contain several layers of network protocols each with own attachment mechanisms, starting from link-layer procedures and IP address resolution, there might be a large number of layers, each requiring some number of hand-shake messages to establish connection. These handshakes could be paralleled. This involves security challenges and requires heuristics to roll back established communication state in case the handshake is terminated on some layer for one reason or another.

- Near term: performance on mobility, hand-offs, and multi-homing using the existing protocols; means for the end hosts and networks to select the best possible access interface; fast hand-offs
- Medium term: isolated network domains; different network characteristics; seamless behaviour; working seamlessly in a mixed IPv4 / IPv6 environment; sub-network mobility.
- Long term: New forms of mobility, e.g., peer-to-peer, infrastructureless networks; new compensation mechanisms.

Compensation, Resource consumption, and Congestion

The current Internet architecture is not well equipped to deal with short-term resource shortages. That is, while the TCP and other related well-behaving protocols courteously back up and reduce their resource consumption in the face of congestion, there is nothing in the architecture itself that enforces such behaviour. Our basic claim is that the lack of resource and congestion control in the Internet is inherently related to the lack of compensation methods.

The Internet was fundamentally built upon the idea of at-least minimally co-operating agents. In the very core of the Internet architecture lies the assumption that if a host does not want to receive traffic, the sending nodes will cease sending. Similarly, almost as close to the core, lies another strong assumption: the hosts and routers will co-operate in getting the maximal amount of traffic through for the maximum number of hosts. These two assumptions are engraved deep into the architecture and the implementations.

Looking from an economic point of view, and assuming selfish rather than co-operating agents, both of the core assumptions start to appear silly. If the senders are not forced to somehow compensate for the resources they use, or at least overuse, there are no way of creating incentives for stopping them.

Looking at the two problems (lack of resource control and lack of congestion control) more closely, we can see that they are actually the same problem, only working on different time scales. The resource control mechanisms attempt to make sure that there are sufficient resources at all times. The congestion control mechanisms attempt to make sure that the available resources are allocated according to some “fairness” or “relative utility” principle in those cases where the demand exceeds the supply.

From a more technical point of view, the challenges include the following:

i) how to satisfy performance requirements of different kinds of new networking applications, such as Internet telephony, video streaming, games, and at the same time implement feasible and fair congestion control in a best effort network,

ii) how to make the congestion and rate control algorithms more reactive to the changing network characteristics, and

iii) how to address other types of resource limitations than available bandwidth and buffer space, such as available power in a battery-powered device.

The possible research work includes:

- Near term: external triggers, make the congestion control more reactive to the sudden changes in the network environment; challenges related to using newly proposed congestion control algorithms.
- Medium term: incentives related to congestion control; ways to encourage correct behaviour; mechanisms to provide increased flexibility to the applications
- Long term: fine-grain adjustments in the flow transmission rates.

Privacy and Attribution

The final two problems on the list of current problems are somewhat different than the ones above. Here, we are interested in preventing bad things from happening, in one hand by imposing restrictions on information flow, and in the other hand by creating explicit incentives for trustworthy behaviour.

The privacy problem can be described from a three different starting points. From the Orwellian point of view, the question is pretty much about freedom of speech and governmental control. The Kafk aesque aspect of privacy focuses on citizen’s ability to retain their autonomy without fear of unfounded litigation or other harassing legal/other action. Thirdly, the economic aspect of privacy relates to the fine balance between socially beneficial differentiated pricing vs. socially harmful price discrimination. From these
three different points of view, it seems a necessity to provide a reasonable base-level of privacy as a built-in feature in future networks.

The flip side of privacy is attribution. Unbounded privacy encourages unwanted side effects, such as rampant advertising (spam). To counter these, increased privacy requires increased accountability.

There are two hard problems related to privacy and attribution: First, how to provide both of them at the same time, and, second, what is the right balance.

**Trust and reputation**

The final problem we consider is the lack of trust and reputation. The original Internet architecture was build with a fairly homogenous, mutually-trusting community in mind. Today the user community is very large and diverse. The hosts cannot be trusted to respect protocol specifications any more, due to prevalence of botnets and other malware. More generally, we postulate that the current network and applications suffer from the lack of standardised, wide spread mechanisms for asserting trust and reputation.

Different trust domains:
- banking, or more generally, different methods for expressing accumulated wealth and liquid currencies
- health care

Retaliation through the legal system; stability of the legal system, co-dependencies

**Major future challenges**

**Information networking**

The network today is no longer used just to convey written or otherwise recorded messages or pass real-time voice or video, but for entertainment, e.g., interactive gaming, and, more generally creating a new kind of social consciousness. The younger generations consider it normal to be "on-line" all the time. More recently, the usage patterns have started towards sharing presence and experiences; i.e., using the network to roughly keep track of the whereabouts of the friends and other acquaintances, allowing new kinds of quite dense social structures to emerge.

At the same time, the proliferation of intelligent, networked devices and search services is gradually making it impractical to identify information by the device hosting them. In many cases it is impartial if the information is found in a given format or another, or where within the global Internet the information can be found from. As can be seen from the usage statistics of peer-to-peer networks, there are strong incentives for both acquiring data and making it available even in the case where the legal property holders of that information would like to inhibit such practises.

From the networking point of view, this new viewpoint imposes a huge challenge. Once again, the whole architecture needs to be rethought. Once again, the concepts of naming, addressing, and routing will need to be completely re-scrutinised, and most probably some completely new concepts and technologies will emerge, similar to the emergence of routing protocols as a result of the first revolution.

On the other hand, the problems of privacy, accountability, trust, and reputation still pertain, though in a slightly different form. For example, since the object of interest is now data instead of connections, the privacy problems will now condense around data creation and consumption instead of connection tracking and eavesdropping. Similarly, while the problems of node trustworthiness and reputation become less obvious (though perhaps no less important), a new set of problems related to data content and its reliability emerge.

**DTN**

Location independence; ability to compute anywhere, ability to utilise parallel computing power

Ability to carry and cache data round in mobile devices

**Usability and changing usage patterns**

XXX Usability and complexity; Martti.

Interplay between the way users take technology to their use and the social structures. New emerging social structures. Dunbar’s number. The role of trust. Transparency.

Given that the Mark Weiser’s vision of transparent, ubiquitous computing turning out not to be one that people want, the complex two-way interactions between users’ behaviour and social structures, including co-evolution of institutions and technology, and the emergence of new patterns from the changing individ-
ual behaviour, we postulate that it is virtually impossible, at this stage, to anticipate what kind of traffic patterns the evolving usage patterns will produce. Consequently, it appears even more important than before to bridge the gap between applications and networking researchers. For the future Internet, the networking people are looking at radical ideas, such as information networking. At the same time, the applications people are largely assuming that the basic form of networking will continue to exist at a host-to-host bases. This disconnect is likely to lead to largely misleading projections about the likely evolution at both fronts.

At a higher abstraction level, the new types of applications and social interactions are likely to impose people and businesses to new types of risks. Hence, in order to facilitate models for risk management and assessment, the insights for new interaction models and resulting traffic patterns need to be enhanced with strategic\(^1\), trust related information and models of how people handle it.

Finally, at a more macro level we anticipate that the changes in this arena will necessarily lead to formation of new business models. The immediate causes are likely to spawn from internalising the resulting reduced transaction cost and the tensions created by the uneven creation of new wealth. In the longer run, it looks also very likely that completely new forms of compensation will emerge, perhaps more suitable for informal or non-institutional social interactions. From an economic point of view this will mean that part of the total economic activity will move from the "dollar markets" to local, social markets. Among other things, this creates new challenges for measuring the total amount of economic activity taking place in the society.

**Autonomy and Resilience**

Despite its urgent need and apparent efficiency benefit, there has been relatively little real life progress towards genuinely autonomous and resilient networks. This realisation creates two distinct problems:

- Understanding the socio-economic factors that have largely hindered the adoption of even those few pieces of autonomy or resilience enhancing technologies that exist.
- Search for new approaches towards building autonomous and resilient networks, taking into account the aforementioned socio-economic factors.

As a first approximation, we posit that ideas and techniques from complex adaptive systems (CAS), which have been applied successfully to explain certain aspects of biological, social and economical phenomena, can also form a partial bases for building autonomously resilient networks. However, in order to reach there we expect that a huge amount of research is needed to better understand the mechanisms behind the emergent socio-economic features facilitated by the current Internet. Only such understanding will allow one even postulate the potential micro-level formations that are needed in order to facilitate the emergence of resiliency at the macro-level. In other words, our initial assumption is that, given the complexity constraints, it is impossible to design networks that are genuinely autonomous and resilient at the same time. Therefore, resilience and genuine autonomy shall be seen as emergent properties, growing from the complex structure of underlying interactions, rather much in an analogous way to a living cell being an autonomous and resilient system.

**Energy consumption**

The energy consumed by high-tech industries and institutions represents an attractive and often untapped opportunity for energy savings. Characterised by large base-loads operating 24 hours a day with energy intensities much larger than typical commercial buildings, high-tech buildings include laboratories, clean-rooms, and data centres. These facilities are essential to various industries important to the national economy. As it relates to data centre operations, establishing measurements and metrics for energy efficiency will guide managers on what results to expect from investing in green technology.

Remedies from cross-layer, link-layer, MAC, and network design. In cross-layer design the fundamental questions include to define the set of information that should be exchanged across protocol layers and how should that information be adapted to, and how should global system constraints and characteristics be factored into the protocol designs at each layer. In link layer design the goal is to the achieve fundamental capacity limits of the channel; i.e., to overcome channel impairments with relatively little energy. This may require new coding techniques, drawing ideas from block codes, convolutional codes, trellis codes, and turbo codes. Closer to the physical, MIMO and multiple-antennas are important. Last but not least, power control is needed to compensate for random channel variations, reduce the transmit power, minimise the probability of link outage, reduce interference to neighbouring nodes, meet hard delay constraints, and prevent buffer overflows. Adaptive resource allocation needs link transmission scheme transmitted power level, symbol transmission rate, constellation size, coding rate/scheme.

At the MAC layer, the main problems include how to divide the spectrum into different channels, how to assign these different channels to different users, or channellisation (time/frequency/code division) and

\(^{1}\) We use the term "strategic information" here in its ethnographic meaning, denoting information that people collect about other people in order to be able to interact socially.
scheduling. In a nutshell, how to organise each node on different channels to best use the available time, frequency, and energy.

From the networking point of view, problems include neighbour discovery and network connectivity: probing, transmission range, adaptive rate, power, and coding; routing: flooding, robust, less routing overhead, waste bandwidth and battery. In energy-constrained routing issues to consider include delay constrains, battery lifetime, and routing efficiency.

**Shifting bottlenecks**

Hugely increasing core capacities

Ever growing difference between (wireless) edge and (optical) core capacity (It would be good to get some text to here from Raimo Kantola’s lab)

New memory technologies.

Hugely growing disk space & caching capacity, again.

End of Moore’s law; moving to multi-core CPUs. Affect on bottlenecks?

**Towards a Vision and Mission**

A backbone for the society. Major effects on the efficiency of the economy. Complete collapse would have dire consequences. Distributed resiliency, ability to function independently from any central components is a necessity.

Creating new wealth. Lower transaction costs, especially search and trust/risk related costs. Support even more diversity of products and services. Agile markets.

**Problems requiring major breakthroughs**

Completely autonomous resiliency; Zero configuration, something called configuration agility. Current way of joining the inter-connection and roaming community requires manual work and a lot of it. If we assume that network edges and operators show up and disappear frequently the process and distribution/removal of ‘contact’ information should be instant. Same applies also for end users that change their subscription owner but retain their identity. Also having the configuration information distributed without a centralised management is definitely something to look for.

Economics of identity, privacy, and attribution. The technical solutions are relatively easy (see below): the hard part is how they will interact with the rest of the system.

**Missing components**

Technologies for expressing and inferring on reputation and trust

Identities with balanced and strengthened privacy and attribution

New types of compensation methods; disentanglement of the functions of money; currencies based on other models besides fractional reserve banking

Structured market places to internalise the most obvious tension points.

**Contributors**

The main editor for the present version was Pekka Nikander. The main technical background thinking was performed by Hannu Flink and Pasi Sarolahl, who provided lots of text about the technical details. Other contributors include Jouni Korhonen, Jussi Kangasharju, Christer Carlsson, and Martti Mäntylä. Some pieces of the text were stolen from a joint paper written by Dirk Trossen, Sasu Tarkoma, Mikko Särelä and yours truly.