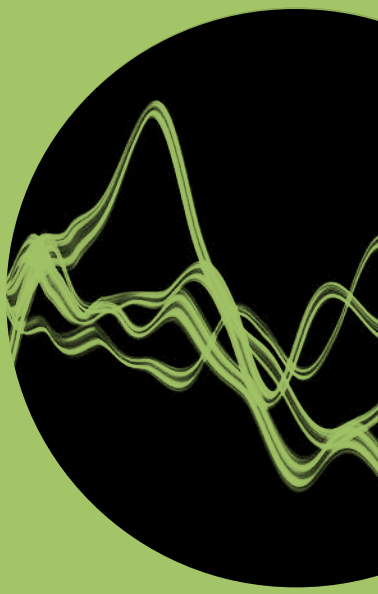
CHARTING THE FUTURE OF INNOVATION | #04·2021

ERICSSON TECHNOLOGY Review



USING AI TO ENSURE ENERGY-EFFICIENT NETWORKS

ERICSSON

Ensuring energy-efficient networks

WITH ARTIFICIAL INTELLIGENCE

Finding ways to make networks more energy efficient without negatively impacting QoE is critical to network operators for both cost and sustainability reasons. To assist in these efforts, we are exploring the potential of using artificial intelligence (AI) techniques to recommend energy-efficient configuration settings for network nodes.

KONSTANTINOS VANDIKAS, HELENE HALLBERG, SELIM ICKIN, CECILIA NYSTRÖM, ERIK SANDERS, OLEG GORBATOV, LACKIS ELEFTHERIADIS

Our estimates indicate that the cost of the energy required to power networks represents between 10-30% of the network operating expenses of a communication service provider (CSP), depending on the specificities of its local energy market. In total, this expenditure adds up to approximately 25 billion USD per year [1].

■ Despite the many energy-efficiency solutions already implemented in mobile networks, energy consumption continues to rise in response to the rapid growth of both network traffic and data volumes. Our research indicates that additional energy-efficiency gains can be achieved by using machine-learning (ML) techniques that enable higher levels of automation.

ML is a type of artificial intelligence in which models learn patterns from data without being explicitly programmed. By recommending configuration settings that can be applied to base stations and other equipment, ML-based techniques make it possible to reduce energy consumption in network elements without impacting QoE.

Access nodes are at the center of our work to improve energy efficiency in networks. An access node (often simply called a node) refers to the relationship between connected user devices and the network elements to which those devices are connected. Access node configuration settings strongly influence node energy consumption and potentially many observable network performance QoS metrics.

As configuration settings rarely change at a

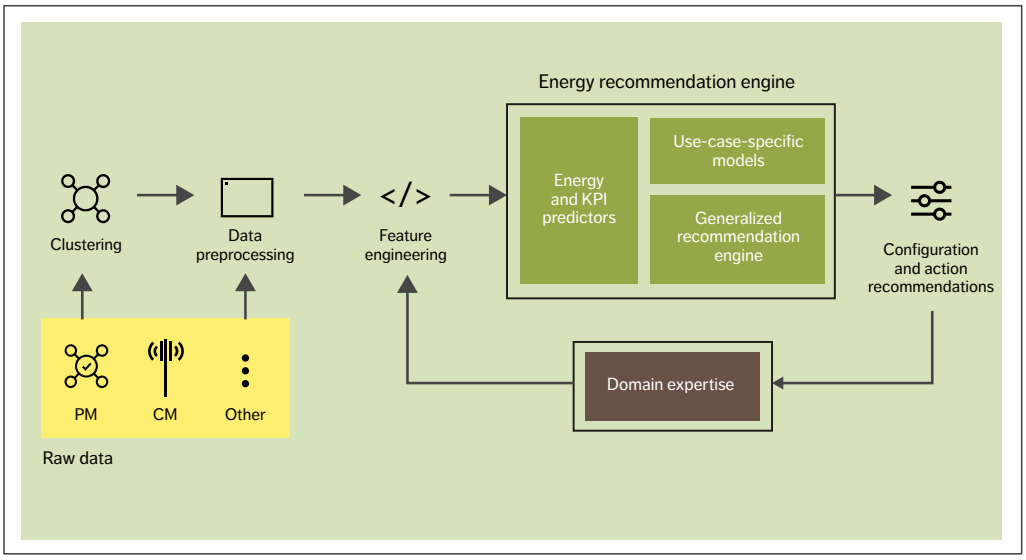


Figure 1 End-to-end energy optimization from power system to node to network

particular node, the ability to regulate node energy consumption requires a mechanism that enables the generation and evaluation of new configuration settings to explore their impact. To avoid generating configurations that may negatively impact existing key performance indicator (KPI) levels, new configurations need to be bounded by different KPIs that constrain the search.

Given that some configuration settings may require more time than others to take effect, there is a need for accurate predictive models that make it possible to foresee when such changes can be applied ahead of time. This would then minimize any potential disruption to the network's operation. An ideal solution would identify as many potential ways to reduce energy consumption on the current functionality of the network elements.

With all of this in mind, we have developed a concept for end-to-end (E2E) energy optimization that encompasses everything from the power system to the nodes to the network level. *Figure 1* illustrates our concept, highlighting the energy recommendation engine that is at its core.

Data set for the energy recommendation engine

All the data used in our work on the energy recommendation engine was collected from a live network. We primarily used performance management (PM) and configuration management (CM) data as measured in the base station, where

Terms and abbreviations

CM – Configuration Management | **CVAE** – Conditional Variational Autoencoder | **DL** – Downlink | **E2E** – End-To-End | **GNN** – Graph Neural Network | **KPI** – Key Performance Indicator | **ML** – Machine Learning | **PM** – Performance Management | **PRB** – Physical Resource Block | **PSU** – Power Supply Unit | **UL** – Uplink

> TO ENSURE ZERO NEGATIVE IMPACT ON QoE, WE USED KPIs TO CONSTRAIN OUR MODELS

energy measurements are already part of the collected dataset within the PM. As a result, there is no need to deploy new hardware to get the data. The radio network performance counter data sets contain observations on cell performance such as the activity count in downlink (DL) and uplink (UL) directions, the utilization in the cell and units.

To ensure zero negative impact on QoE, we used KPIs to constrain our models. While it is technically possible to include additional and/or alternate KPIs based on operator preferences, we selected these five based on how they affect energy consumption:

1. Number of connection attempts to a cell
2. Average number of users in a cell
3. Throughput
4. Latency
5. Interference.

When telecom networks are installed, they are typically configured with certain parameters such as the number of cells and the hardware unit types (indicating frequency bands) and so on. Possible reasons for a reconfiguration could be a problem such as a software issue or a hardware failure after the installation of new parts that may come from different vendors. Over time, subtle changes and different tuning may lead to different energy consumption levels, where, for the same amount of traffic, this may in some cases be positive (less energy consumption) while in others it is negative (more energy consumption).

The CM data set that we used consists of hundreds of configuration attributes of the sector, including the settings of each radio cell (such as frequency in DL and UL directions) and installed hardware types. We used this information to be able to recommend multiple configuration changes at once, rather than focusing on one at a time.

The output of the energy recommendation engine consists of a set of configuration attribute changes for a corresponding node. The output captures the interplay between different nodes and configurations rather than focusing on isolated fine tuning on a per-node level, which may have an effect on other nodes.

Methodologies used in the energy recommendation engine

A content-based recommendation model requires a good representation of the configuration settings in the embedded (also known as latent) space. Such a representation can be hard to obtain manually due to the high number of configuration attributes.

Conditional variational autoencoders [2] (CVAEs) and graph neural networks [3] (GNNs) are two of the most suitable and complementary techniques for our purposes, as they are both fueled by the success of neural networks. While a CVAE is generative and adversarial, helping to explore large spaces in bounded conditions, a GNN can act as a critic (or discriminator) that can suppress abnormal recommendations, especially when the recommendation model diverges too much.

In addition to GNNs and CVAEs, we also used conditions to confine the new configuration settings under specific KPIs that should not be broken while new configuration settings are created. Such conditions may originate from domain expert engineers, while others can be universal or specific to the network that is being examined.

Graph neural networks

GNNs offer a straightforward way to learn from relational data. We use them – and graph convolution in particular – in two ways in this project: to generate conservative recommendations for energy efficiency based on historical information and to generate multi-site predictive models for different KPIs. Multi-site predictive models are essentially enhanced forecasting models that help operators predict performance based on KPIs that serve as measures of how well a network is behaving.

One of the weaknesses of conventional forecasting techniques such as long-short term memory is the inability to account for the spatio/temporal relationship that exists between nodes. This is problematic because nodes are positioned strategically in various parts of urban or rural spaces and configured accordingly to serve the requests made by user equipment in their vicinity. Information about their spatio/temporal relationships is highly relevant in forecasting.

We have studied the possibility of combining graph convolution with regular 2D convolution. This approach combines two inputs: the adjacency matrix that represents the network's topology, and the time series of each node for a different KPI. The network's topology is constructed using geographical information per node. This information is then represented as an adjacency matrix, which captures the temporal aspects of each node.

Time-series information per node is preprocessed to produce the corresponding input and output predictive window. In this case, we use 12 hours and learn to predict the next 12. Graph convolution and regular 2D convolution is interleaved to build a combination of the two. Our results indicate that this approach increases model performance in multi-site prediction.

The second way that we have used GNNs in this project is to generate conservative recommendations for energy efficiency based on historical information – that is, we used GNNs to create a recommendation engine.

To understand these changes, we represent the relationship between different nodes in a telecommunication network and their configuration sets as a graph. In graph theory, a graph is a structure that contains a set of objects (or nodes) that are connected to each other through links. A link between two nodes means that the two are associated. In this context, we consider a heterogeneous graph, as the objects that we connect are of different types. More specifically, we associate nodes with configuration sets, and the association is represented by a link between them.

In addition, each link is labeled based on the efficiency of that association from an energy perspective as well as on the number of connected subscribers. For each of these elements of the graph we learn a representation of that, driven by its features such as the hardware installed or the different types of parameters in each configuration set.

As a result, the problem is transformed to a link-attribute prediction problem, where we train a model that learns to predict how energy efficient that association was. Even though this system is not capable of generating new configuration sets, we see that it is capable of matchmaking similar nodes with potentially similar configuration sets that can have a lower energy impact than the one currently used.

Conditional variational autoencoder

The high-level definition of a generative model infers generalized distribution of observation features that are confined by the constraint conditions. The CVAE generative model-based recommendation engine consists of multiple components including an encoder, decoder, prediction model for the target KPIs and a prediction model for the energy-consumption target.

The encoder model compresses the representation of the raw CM dataset into a low dimensional matrix, which is referred to as latent space. Latent space consists of points, where each point in a 2D latent space represents a full CM configuration setting that is constrained with an energy consumption and a KPI value.

Due to the nature of the CVAE, the CM configurations with the same energy and KPI constraint categories are located close to each other in the latent space. The decoder reconstructs the complete configuration file, which could consist of many CM attributes (configuration parameters and settings), from the embedding representation in the latent space. The latent variables that represent the targeted KPI and energy-consumption levels are drawn by uniform random selection from the many that represent the same category of the targeted KPI and energy levels. The decoder then uses the input to generate new configuration attributes.

In deployment, we provide constraints as input to the decoder of the CVAE model bounded by the

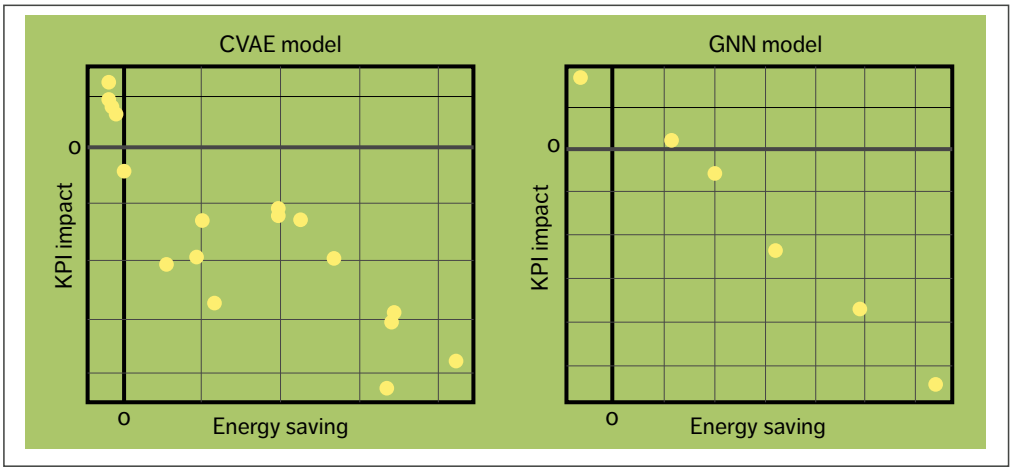


Figure 2 Energy savings and KPI impact according to the CVAE model (left) and the GNN model (right)

required KPI and energy values. We can use multiple constraints and they can be customer specific. The constrained energy value is set slightly less than the predicted energy consumption, as an energy-efficient synthetic CM file is expected as the output of the generative model.

At the same time, we aim to sustain the KPI value. The predicted KPI value at a network node is therefore given as input to the decoder. There are two different prediction models for KPI and energy models:

1. A CM-based prediction model that only uses CM attributes in tabular form as input
2. A PM-based prediction model that only uses PM counters in time-series form as input.

The time-series PM-based prediction models perform 24 hours in advance to give enough time to deploy the desired configuration to a corresponding network element. Prediction results from CM and PM models for energy consumption and KPI (connected users) values performed well. This is important for the accuracy of the generative model output, as the outputs of PM-based models are used as inputs to the CVAE generative model together with the selected latent variables mentioned earlier.

It is important to quantify the amount of energy saved with the generated CM configuration set compared with the existing planned configuration set. For that reason, we used a pretrained CM-based energy model, which only takes CM attributes as input features, and this model predicts energy consumption using purely CM attribute value combinations. This means it can contribute to estimating a mean base energy consumption value given a configuration.

First, the configuration generated by the recommendation model using the actual predicted energy as a constraint is given as input to a pretrained CM-based energy model. Next, the configuration generated by the recommendation model using a value that is lower than the predicted energy as a constraint is given as input to the same CM-based energy model. Finally, the difference between the two energy predictions is computed, and an indicative potential saving is quantified.

We repeat the same steps on a pretrained CM-based KPI model to quantify the KPI impact. This allows us to obtain a KPI versus energy trade-off curve as shown in *Figure 2*. In general, these curves tend to show that higher energy savings yield a higher negative impact on the KPI.

Comparison of the conditional variational autoencoder and graph neural network

A comparison of the CVAE and GNN reveals that, given its generative nature, a CVAE produces new configuration settings. In contrast, a GNN only identifies potential energy savings as marked by a rating function. In this case, the rating function contains six distinct categories, each represented by a dot on the right in Figure 2. (While a GNN is non-generative in the context of our research, the literature indicates that it can also be used in a generative context.)

The comparison also reveals that the efficacy of the CVAE-based recommendation model is dependent on the accuracy of multiple CM- and PM-based KPI and energy prediction models, which means that all the predictive models need to be accurate simultaneously. As a good side effect, this modular structure potentially makes it easy to troubleshoot during the maintenance of the model performance.

Meanwhile, in the case of GNNs, there is only one graph-based model and it is limited by the number of different configurations that are available or have been applied to that network. The use of a single model simplifies its maintenance process.

As both CVAEs and GNNs are predictive recommendation models, it is possible to omit any recommendations that are predicted to impact network performance and QoE in advance. Both models yielded configuration recommendations that are predicted to achieve up to 10% energy savings when applied.

Use-case examples

To further increase the efficacy of our energy recommendation engine, we enhanced its ability to improve energy efficiency by applying recommendations in two specific use cases: radio signal interference detection and PSU load utilization, as shown in *Figure 3*. Both use cases follow the same three steps highlighted in the middle of the figure:

1. Identification (localization) of nodes with the potential for improvement
2. Modeling of the selected nodes to understand their behavior and predict their states
3. Actions to improve energy efficiency (implementation).

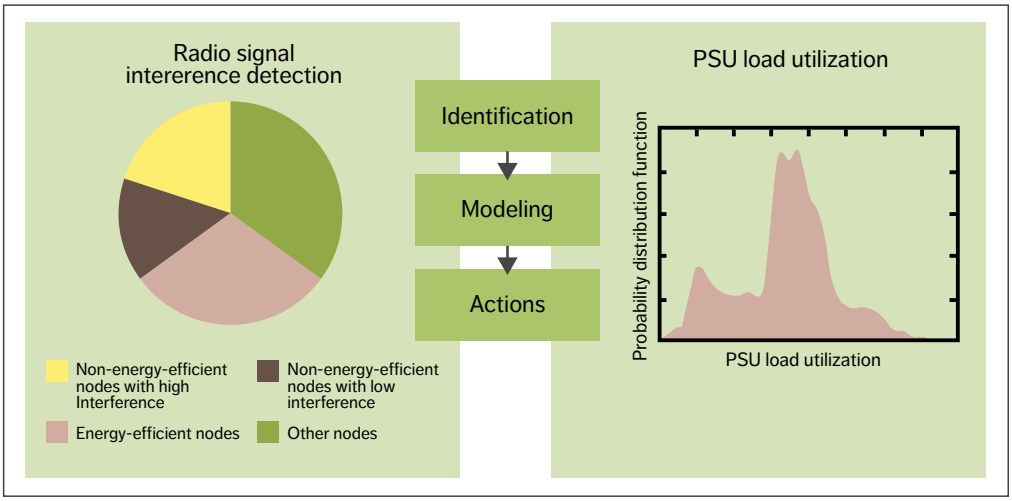


Figure 3 Two use-case examples – radio signal interference detection and PSU load utilization

Use case No. 1: radio signal interference detection

Interference is created by a range of factors such as changes to the environment or having too many users in the same cell, especially if they are located close to the cell edge. It is also possible for cells to interfere in frequency with each other. Interference has a significant impact not only on QoS but also on power consumption. While current radio equipment is designed to handle interference in a way that avoids unwanted emissions and variable techniques are used to limit the interference in the network, radio signal interference has proven to be a formidable challenge to overcome completely.

> A HOLISTIC APPROACH IS THE BEST WAY TO ACHIEVE OVERALL ENERGY SAVINGS

By introducing clusters and dividing the problem into sub-problems, our energy recommendation engine makes it possible to identify the nodes that have high interference and low energy efficiency. The left side of Figure 3 shows a breakdown of radio signal interference into four categories: non-energy-efficient nodes with high interference, non-energy-efficient nodes with low interference, energy-efficient nodes and other nodes. The possibility of improving the nodes classified as other nodes will be addressed in the future.

By modeling the network traffic load, consumed energy and interference-related KPIs, we can save energy by recommending action on the interference cells (such as locking them) to avoid the high energy-consumption state. The cell interference parameters are therefore unique and connected to the network activity and consumed energy.

Use case No. 2: PSU load utilization

A radio base station consists of several PSUs supplying power to the radio units. It is not uncommon for PSUs in radio base stations to be underutilized, as illustrated by the graph on the right in Figure 3. Active but underutilized PSUs may not be working efficiently within the operational range of the unit and consume more power that necessary due to power dissipation. It should be noted that PSU efficiency depends on the load.

To identify nodes for PSU load utilization improvement, we clustered them according to the PSU load, the number of PSUs, relation to the radio network activity (such as the relative radio resource usage), the number of active users, PRB utilization and the number of connected users. We focused on PSUs with less than 50% utilization. Our research show it is possible to propose dynamic power-supply control such as putting underutilized PSUs in sleep mode or turning them off.

PSU efficiency is highly dependent on the load that is applied on the PSU output. Setting one of several PSUs in a system to sleep mode enables savings of 1%. At the same time, the improved utilization and operational efficiency of the PSUs that remain active can provide an additional 1% in savings.

Conclusion

One of our core goals at Ericsson is to continuously improve the energy efficiency of networks. A holistic energy optimization approach is the best way to achieve overall energy savings because it ensures that improvements achieved at one level are not canceled out by increased energy use at another. Our end-to-end (E2E) energy optimization concept is driven by an energy recommendation engine that is powered by artificial intelligence. This solution has great potential for automation, with the help of specific interfaces that can tune the nodes directly without human intervention. It can be fully software based, without the need for additional hardware.

The energy recommendation engine analyzes relevant data to figure out how node configurations can be fine-tuned to reduce energy consumption without impacting QoE. Our research indicates that the total E2E efficiency gains (including radio configurations) generated by our approach can be

up to 10% for radio cells and up to 2% for PSU optimization.

On top of building a generalized energy recommendation engine, we are also developing use-case-specific recommendations for different challenges that can be onboarded to the generalized engine at a later stage. In the two use cases we have studied so far, we used predictive models to find the cases where PSUs are underutilized and to detect interference that may cause unnecessary energy usage.

> USE-CASE-SPECIFIC RECOMMENDATIONS CAN BE ONBOARDED TO THE GENERALIZED ENGINE AT A LATER STAGE

References

1. **Ericsson, Network energy performance, available at:** *https://www.ericsson.com/en/about-us/sustainability-and-corporate-responsibility/environment/product-energy-performance*
2. **Advances in Neural Information Processing Systems (NIPS 2015), Learning Structured Output Representation using Deep Conditional Generative Models, Sohn, K; Lee, H; Yan, X, available at:** *https://proceedings.neurips.cc/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html*
3. **IEEE Transactions on Neural Network Systems (2021), A Comprehensive Survey on Graph Neural Networks, Wu et al., available at:** *https://ieeexplore.ieee.org/abstract/document/9046288*

Further reading

- **Ericsson, AI operations and optimization, available at:** *https://www.ericsson.com/en/ai/operations*
- **Ericsson, AI in networks, available at:** *https://www.ericsson.com/en/ai-and-automation*
- **Ericsson, Energy Infrastructure Operations, available at:** *https://www.ericsson.com/en/managed-services/energy-infrastructure-operations*

THE AUTHORS

Konstantinos Vandikas

◆ is a principal researcher at Ericsson Research whose work focuses on the intersection between distributed systems and AI. He has been at Ericsson Research since 2007, actively evolving research concepts from inception to commercialization. Vandikas has 23 granted patents and more than 70 patent applications, and is the author or coauthor of more than 20 scientific publications. He holds a Ph.D. in computer science from RWTH Aachen University, Germany.

Helene Hallberg

◆ is a senior specialist in energy efficiency radio systems at Ericsson. She joined Ericsson in 1988, first working on the engineering of power and backup systems for fixed and mobile telecom equipment. Hallberg is active in energy-related regulatory discussions and standardization activity, and has filed patents in the energy-efficiency area.

Selim Ickin

◆ joined Ericsson in 2014 and works as a senior specialist in AI at Ericsson Research. His current research interests are distributed ML and intelligent software prototyping. Ickin has developed numerous data-driven ML solutions in diverse domains. He has contributed to numerous international conferences, written several journal articles, and holds patents in various subareas of ML within the scope of mobile networks. Ickin holds a Ph.D. in computing from the Blekinge Institute of Technology, Sweden.

Cecilia Nyström

◆ is a program manager for data and analytics within Business Area Managed Services. She joined Ericsson in 2016, and in her current role she is primarily focused on data strategy and the development of new AI solutions. Nyström holds an M.Sc. in engineering physics from KTH Royal Institute of Technology, Sweden.

Erik Sanders

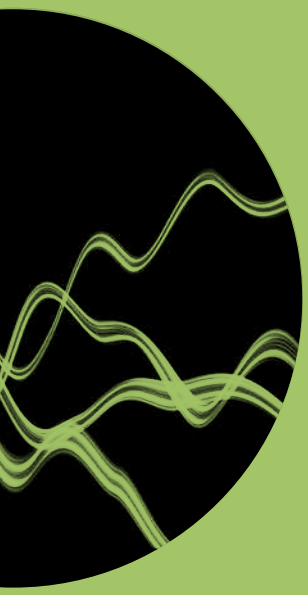
◆ is a product manager for AI and automation within Business Area Managed Services. He joined Ericsson in 2005 as an engineer in 3G RAN and continued with hardware development for radio base stations. In his current role, he drives innovation programs for Business Area Managed Services in the ML and reasoning area. Sanders holds an M.Sc. in mobile communication from Linköping University, Sweden.

Oleg Gorbatov

◆ joined Ericsson Research in 2020 as a senior researcher. His research focuses on modeling and optimization of complex systems. Gorbatov holds a Ph.D. from KTH Royal Institute of Technology, Sweden.

Lackis Eleftheriadis

◆ joined Ericsson in 1998 and is currently a senior specialist in sustainable AI operations within Ericsson Research. Prior to his current role, he had been involved in the development of power products, including functionality for radio access and site infrastructure. Along with several patents in his area of AI, power and energy efficiency, Eleftheriadis holds an M.Sc. in electrical engineering from Uppsala University, Sweden.



ISSN 0014-0171
284 23- 3357 | Uen

Ericsson
SE-164 83 Stockholm, Sweden
Phone: +46 10 719 0000