FOUNDATION OF DIGITAL BUSINESS

Surojit Mookherjee

Vinod Gupta School of Management

Indian Institute of Technology Kharagpur

Week 08

Lecture 40

Lecture 40: Sustainable Digital Technologies

Good morning. The concluding session in this course Foundation of Digital Business, last module was Future Ready Organizations and I will be talking about sustainable digital technologies. I thought I should end my course with this because whatever we are doing with technology it has an impact on the environment and we know the problem we are going to face with rising carbon footprint, rising temperature.So, the discussion is there in various forums in all technology field, in all science be it physics, chemistry or society, climate changes, the impact, the floods we are having, the cyclones, the tornadoes, the earthquakes and everything is we are informed lot of major impacts or changes which are affecting our daily lives.

And of course, the economic environment so to say since we are talking about business, so we will stick to business and see what are the impacts of technology which are going on. And they are kind of unavoidable because to do something if nothing comes free, so like everything has a price for it. So, if I have to use technology, I have to use power for example. So, if I have to use power I have to consume fossil fuel for example, I have to travel we are increasing tourism travel etcetera which is an industry. So, the every industry tries to grow.

So, the tourism industry wants to grow. So, you want to travel. So, you spend more money on fuels. Then we construct lot of things in the hills and mountains etc. Then we have landslides and it can also may be tree or we do not know earthquakes or dams or water supply etc.

So, various things can affect the way nature behaves. And finally, we are completely dependent on nature for our survival.So, something majorly goes wrong with the nature it

will be very difficult for us to survive and the way the temperatures are going up if it grows increases by may be another 5-6 degrees, then we will have to something do something very drastic otherwise it will not be there. So, anyway coming back to our real world real-time energy efficient from this AI or digital technology perspective we will stick to energy efficient algorithms, databases, energy efficient data centers, clean AI etc. and see what are the impacts and as responsible citizens and as responsible professionals.

The main idea why I am telling you this is because all of us know about this is nothing new I will be talking about is just to be conscious that whatever we are doing, whatever we are using technology for our health, we must be aware that these are all consuming energy.So, can I for example, minimize that? I do not want to eliminate that. For example, simple things like we do when we go out of the room, we switch off the lights, fans, air conditioners, because air conditioners are large consumers of energy. So, do we do that?

Many times we find people not doing it. So, can we educate people to that level that yes, whenever you are not using something, please switch it off, do not consume electricity unnecessarily. To give the perspective, the exponential for in our AI world, technology world or the world of internet, the exponential growth in traffic, data traffic. So, if you see internet video for example, a peer to peer sharing, wireless data, wireless voice. So, these are just some numbers.

The chart is just to show how things are doubling almost every 2 years: 40 percent per year, 30 times in 10 years, 1000 times in 20 years, etcetera. So, these numbers are just to tell you that things are increasing at a very fast rate. Now, the environmental impact of digitalization: suppose we use something and what it costs to make that thing again from an energy perspective. So, using digital systems consumes already today more energy than producing those digital systems, and it is rapidly growing—the share is growing. So, here, if you see on the right-hand side, it is manufacturing.

So, here, the share—not market share, the energy share—is 45 percent. And for use, it is 55 percent. So, of the total 100 energy spent, 45 was consumed during the manufacturing of those devices, and 55 is being spent when using those devices. So, devices are like computers, laptops, etcetera; then you have data centers and, of course, the network. The energy consumption of training an AI model—this is very interesting, and you should make a note of this.

One AI model training is almost equivalent to 300 round-trip flights between New York and San Francisco—that is almost a 4-hour flight. So, 4, 4, 8 round-trip is 8 hours of flight; 300 of those—the amount of energy they consume, the jet fuel— is what is consumed in training one AI model—just imagine. So, when we talk about this AI model development—this foundational model, this ChatGPT, Gen AI, Gemini, Gork—we talk about those very casually. But you must also think that to develop each of these models, how much energy was spent.

This session we only talk about energy, not about the manpower, the skill, computational skill, etcetera, etcetera or the equipments, just the amount of energy spent in running those data centers. What is equal to almost 5 car life cycles in US? The life cycle of a car in US could be may be 10 years or whatever, unlike India which is much longer and they travel a lot, drive a car, everything they do car. So, the amount of fuel they consume in a car life cycle 10 years, 5 or such cars.We measure it in pounds of CO2 equivalent carbon dioxide generated equivalent because that is carbon dioxide is what is creating the problem of the global energy or the temperature, rising temperature, global warming.

The round trip flight between New York that is the number numerical data is 1984 pounds of carbon CO2 equivalent. Human life average about 11,000 pounds and US energy they spend much more energy than we do 36,000 pounds of CO2 equivalent. A car the lifetime 126,000 and a transformer model with 213 million parameters, with neural architecture such that this is the AI model how they are trained in the model, it consumes 626,155 pounds of CO2. So, it has training you can see has enormous energy requirement.

Even before the corona crisis the total global carbon emissions of digital technology was Now, it is of course, will be much more.Energy efficient data centers, who are the main guzzlers of energy in this technology world is the data centers, because for everything you do you need a server and the servers are residing in data centers. So, sustainable data centers are just more than just being energy efficient, they are also impacting many other things I will talk about that little bit gradually. And this just as a reference there are 17 UN sustainable development goals by United Nations.

You can have no poverty, zero hunger, good health and well-being, quality education, gender equality, clean water, sanitation, etc. Affordable and clean energy that is number 7, climate action that is number 13 and other things there.So, 17 you can I mean refer to it is now quite well known, but this is what we are now referring to how digital technology

the world of AI is impacting affecting the climate situation the climate action that is number 13. Now digitalization in numbers the traffic volume internet every minute of the day this is the 2020 figure I could not get a later figure from the day, but does not matter

here is not important it just shows indicates what is the huge internet traffic all of us are generating.Take Netflix 404,444 users they stream, so many hours of video. If you take YouTube, the users upload 500 hours of video every minute, every minute we are uploading video files to the extent of 500 hours of viewing time. Facebook users share 150,000 messages every minute across the world of course, this is Amazon is shipping 6600 packages every minute and the number of video calls we make, video or voice calls 1 million, 1.3 or 1.4 million video calls that is 13 lakhs, 13,80,000 etcetera

video audio calls we are making every minute and why do you have to talk so much?If you ask, you see people unnecessarily spending time talking. Or every day morning you get good morning from so many of your friends in the WhatsApp group, every day good morning and everybody sending good morning to everybody. So, just think the spiraling effect. Compounding effect. So, many billions of good morning messages going every day I mean what value does it add every day you get 50 in your WhatsApp inbox good morning, good morning,

good morning some picture flower etcetera, smiling face.What is the value that is adding? Why do we need this? So, sometimes you should sit back and think the whole purpose of showing this. Do we really need to do so much? And at what cost?

So, think for the cost angle then only you will think backwards. First think what is the damage or cost and then think backwards. Because for us it is not pay fixed cost, we have the network or the data card fees fixed is unlimited. So, we pay 600 rupees a month and then we got the entire freedom to do keep calling 24 hours a day or using data to whatever level. Data centers are becoming more climate friendly the average greenhouse gas emissions of data centers in Europe are already falling significantly.

So, that is kind of a good news that technology because technologists and scientists always trying to solve problems. So, this is a problem.This is here to stay because a data centers are required, energy will be required, number of data centers will keep increasing because we are increasing the use more and more of AI models etcetera, but can we then alternately make them more energy efficient. So, that is what everybody tries to do and that is what we are doing because we are intelligent people and we are smart people. By

2030 if this is say this western Europe which was bit increasing, but then has decreased again this is a prediction forecast and then different regions Scandinavia, southern,

rest of northern Europe, eastern Europe etcetera.So, overall you see it is decreasing the energy consumptions in the data center. The energy requirements are servers and data centers in Germany. Figure from Germany, but it is true for anywhere the total energy requirement is obviously, going up. Because more and more IT services they increase the more and more data centers will depend.

So, total energy requirement gradually overall it will go up. The projected rise in energy consumptions of data centers till 2030, I got this chart from Gort, they just made a query from and see what it does. Figure 2030 can you tell me a forecast, give me a forecast and this it referred to an international energy agency and some other agencies data from them took and collated for me. It shows that in 2024 it is about 414 terawatt hour TWH and it goes up to 900 something in 2030 terawatt hour. This global data center energy consumption.

I wanted this figure. So, got it from there. Thanks to Gork global data energy data center energy consumption terawatt hours. The energy efficiency of data center is increasing very significantly. We are making them more efficient.

In terms of computing and storage capacity the energy requirements the data centers have been reduced by a factor of 6 to 12. Cloud based solutions enable an additional boost in efficiency. So, if you know a manager has to go for a new technology, the question is do you have the data center of your own under your control in your premises or you will use a cloud service provider.

If you think from energy go for the cloud solution plus of course, there are other advantages too. So, now this session we just talk about the energy. For energy perspective yes cloud is better than your on-premise data center. A data center infrastructure for air conditioning for example, power supply, fire protection, etc. become significantly more energy efficient.

So, data centers are huge heat generators. So, they have to be air conditioned, very strong air conditioning, power supply, fire protection. All of these need energy, but then significantly improving the efficiency. Some technological potentials which are for improving further the energy usage of data centers, efficient algorithms and optimized

software program. You are basically making them work less for delivering the same output.

New hardware solutions end Moore's Law. Moore's Law is about doubling the number of transistors on a semiconductor chip. People have seen this law in effect since the 1960s, but now they say we need something more. We need more drastic changes in chip design and hardware manufacturing to improve energy efficiency. Everyone in the chip industry focuses on energy consumption, aside from its primary function. The main concern is how much energy it uses.

Because millions of chips are used in servers and data centers. Total energy generation becomes the big problem. Better utilization through management tools. Optimization of cooling, air conditioning, etc. That is why many companies build data centers in cold countries like Iceland, the Nordics, Norway, Finland, and Sweden to use natural cooling and reduce air conditioning and water cooling needs.

Using renewable energy, sector coupling, and waste heat. Renewable energy—many still use solar and wind to power data centers. Companies like Google, for example, are making their data centers nearly zero-emission. That is the goal. Now, how to save energy in daily life—what we can do in small ways.

My digital life consume electricity as in the we have to consider the production part manufacturing these devices. So, can I use less devices or can I change my phones less regularly. So, instead of using one year every year changing etc. Can I do it less?

So, I consume more hardware daily usage talking about those phone calls and videos and sharing good morning wishes etc. That is same true for internet traffic and of course, data center if I am using in data centers etc. This is again I am referring to a study done by HPI student sustainability club, HPI is subsidiary of Hewlett Packard. I have used some in this study some taken some pictures from there. How does my digital life consume electricity?

So, if you take this overall my usage. Video streaming itself 60.6 percent, web browsing, gaming, social network, file sharing, marketplace, cloud, messaging, some security services and audio streaming is this, the lowest one. The internet is responsible for 3.7 percent of total global CO2 emissions and video streaming alone accounts for 61 percent of those emissions. So, does this give you a message? 60.6 percent energy consumption is due to video streaming.

Do I need to share literally and seriously those video files to in a group for example? That is going out to suppose I have a group of 60, if I send a video it is going out to 60 devices. Is it justified? Is it required?

We are so casual about casually forwarding videos to group, any size 100 member, 60 member, 200 member, we do not even think, but this should make you think. Data transmission through the internet emits 1.6 billion tons of greenhouse gas every year. The communications industry will represent 20 percent of all the world's electricity consumption by 2025. And more than 50 million tons of e-waste were produced in 2019 alone, a number that is expected to rise by 8 percent each year.

E-waste is all our used phones, computers, batteries, chargers what not. What we can do? Less streaming specially videos, this should be red bold pink size text, I will do that change next. Download files you use regularly, just download them in your device, do not every time go and stream. Streaming music without video, so avoid those video streaming, do the audio streaming it consumes much less bandwidth and energy.

Set lower resolution when streaming audio and video, you can do that they give options you want to do this video at so much maybe etcetera. Choose the one with the lower resolution because I mean what you see just see and then you delete or forget about it 2-3 minutes that is all. If possible do not use a webcam during video conference that is stopping the video unnecessarily if the audio does well I mean why do you need to be seen we ultimately are just talking discussing. Most of the time if you have to present something of course, generally reduce the garbage data, delete unnecessarily email, subscribe newsletter, unsubscribe newsletters. We go on subscribing to many newsletters because it is free and those comes and we hardly see do not even see finally delete, but each traffic is consuming energy.

If I do not read them why do I so, I should unsubscribe that was clean up your keep your mailbox clean and uncluttered. There will be n number of mails coming in and which you do not read and your regular useful mails will get lost in that cluttering. So, better manage your emails might well as clean it by unsubscribing to newsletters or any other things which you do not need and it also saves energy. Giving the examples of these data centers operating data centers, now Google has become CO2 neutral since 2007 and there is no net emissions by 2030. Ecosia already has a negative CO2 balance.

Amazon currently uses 42 percent renewable energy sources, goal is to use 100 percent renewable energy sources by 2025, their data centers and goal is no net emission by

2040. Netflix aims to compensate energy usage. So, these are the big players, these are the commitments. Talking about sustainable technology. Technical profile and value sustainable technology is a framework of digital solutions that can be used to enable ESG, environment social and governance outcomes.

Key actions: increase the energy and material efficiency of IT infrastructure and workplace services—sustainable IT. So, this is what we call sustainable IT. So, can we make the material the equipments more energy-efficient during production and use? When IT infrastructure is a huge backbone, we use a lot of servers, equipment, network devices, etcetera. Prioritize technology investments based on the sustainability issues most

material to your enterprise strategy. So, from a sustainability perspective, which technologies will be more useful to my application organization? So, take it from that angle. Some of the examples you can consider suppose you are the CIO. And you are making decisions on technology investment with sustainability issues in mind or as a focus—what would you do? Options?

To raise utilization rates of shared resource and reduce environmental impact. So, when it is cloud, it is a shared resource. You do not just send data—you own the whole thing. Maybe you are using only 30 percent of it, but it is consuming power all the same, maintaining that full capacity. And you cannot do anything extra about that balance—whatever spare capacity you have. Enterprise greenhouse gas emissions management software facilitates the collection, analytics, and reporting of past, present, and future emissions data.

So, can I log all my emissions data and then find out from the insights what I can do to keep it under control or reduce it etc. The first thing I need is to log the data how much I am actually consuming. Supplier sustainability applications to track the ESG performance of my vendors, my third party, my stakeholders. So, I am considering the whole ecosystem not only my organization, but I have got a host of suppliers, they are supplying many things which for my use for my production goes into production. So, I should be also responsible for the way they are also consuming energy or in their ESG outcomes.

Supply chain can be used things like blockchain technology to protect verify and trace transactions for example, to ensure ethical sourcing that is also another part of the sustainability dimension. Making AI systems more energy efficient, I just got some few

ideas again I refer to use grok for getting it quickly for me. So, software and algorithmic improvements. Model compression, efficient model architecture, sparse computing. These are the things which we can do from the software perspective, how we can make it more efficient, so that it consumes less energy when we are using that software.

We can do it a compression, model compression, efficient architecture and sparse computing. It can use minimum steps to compute.If you look at it from a system level optimization, renewable energy is one of the easy option integration, dynamic resource allocation that can I make my AI systems work during off hours when the energy demand is low not peak, say at the end, I mean midnight for example. So, that is one time even also in many countries they give you cheaper rates of power when you are using it half hours. Can I make them run during night doing all the analytics work.

Cooling efficiency yes of course, for the data centers how to keep them cool because for data centers the biggest challenge is energy management. How to take care of the heat generated by all those powerful chips. The more powerful the chips are the more heat they generate because more transactions are happening.When we talk about supercomputers when we talk about petaflops per second etcetera means those chips are doing calculations at a extremely high speed that is why they are supercomputers or that is why they are more powerful chips. And when you are doing things very fast you are also burning energy.

Edge computing, edge is another type of cloud computing only thing that it is a smaller device and nearer to your place. So, you have latency etcetera factor much lower compared to a cloud. So, these are being used in various applications like one is of course, the autonomous car because you need to have the response very fast. Because, if you see something senses something and then it has to take an action.That difference between it is going to the sensing thing and the action that signal communication has to perhaps between your sensor to the cloud and back to your computer or the car computer has to be done at very fast time.

May be a second or less than a second of time you are talking about. That is where we use things called edge computing which is much closer physically proximity. The same is true for when you do some remote surgeries, robotic surgeries etcetera, there the surgeon sitting somewhere and the robot is doing the surgery and the operation and the sensors, the cameras and other thing devices sending signals to the doctor in his system and then

he does the whatever action he has to do.Again that latency etcetera. These are the places where you can use edge computing, edge computing of a smaller.

They can use much lower category of chips not so powerful no high speed etcetera. In essence they burn much consume much lower energy and what it says here is that it can cut energy usage by up to 90 percent for localized tasks not large scale, big scale, but small localized tasks.So, why do not we use an edge computer for that. Emerging techniques federated learning train models across decentralized devices. So, can I split it up?

All do it at a central system central computer can I break it up into smaller components and do it in different devices. Reducing centralized data center compute needs this can save up to 50-70 percent energy for distributed applications for studies from Google, energy aware algorithms. Incorporate the energy consumption as a training objective using matrix like carbon intensity etcetera to guide the model optimization. When you are training the model, have one of the algorithm a part of the overall thing is the energy consumption, how much energy it is consumed.

Can the model of course, this can take some decision to consume less energy. Of course, the hardware part you can do many things like specialized AI chips which consumes that that is one of the major area of research for any chip manufacturer is how to build chips faster obviously, better chips, but consuming lower powers.Let us conclude. Low precision computing employ reduced precision formats. Many applications you do not need such high power chips.

So, can we have that type of system and employ reduced precision formats, which will consume less energy? And, of course, those chips should also be less costly and cheaper. Similarly, energy-efficient architecture goes because the architecture is the main thing that decides how the chip will function. So, if you make it more energy-efficient, it means you are basically making a chip which uses less energy. Training and data efficiency: apply techniques like transfer learning, few-shot learning, or self-supervised learning to reduce the training data requirement.

Lowering energy-intensive training cycles, because training the model, as I have told earlier, consumes very high energy. Training one model takes the equivalent of about 300 return flights between New York and San Francisco—an 8-hour flight, 300 times. So, techniques like transfer learning and few-shot learning are, of course, part of Gen AI tools, technologies, prompt engineering techniques, and how to train Gen AI models. Pre-

trained models like BERT can be fine-tuned with 10x less energy than training from scratch. So, if you have to develop a foundation model, use one that is already pre-trained; do not try to build everything on your own.

Because, training that developing that foundation model will consume lot of energy lot of training effort.So, why do that use people have already come up with many foundation models use one of those and then fine tuning it because fine tuning will consume much less energy fine tune with whatever data you want to fine tune. Curriculum learning: train models progressively on easier-to-harder data, reducing unnecessary computations and achieving even some 25 percent energy savings, as shown in recent research. So, the way you train can also save energy. Start with the easy steps, then go to complex ones—do not try to dump the whole complex thing right from the beginning, or the computer will struggle. Similarly, batch size optimization: use smaller batch sizes or gradient accumulation to reduce memory demands and energy use without sacrificing accuracy, potentially saving 50%.

So, these are all the management techniques you can employ while training your AI systems. To make AI systems more efficient, some of the best business practices could be optimizing model selection and design, and choosing efficient models. As a business manager or CIO, you should implement model compression and adopt pre-trained models. Invest in energy-efficient hardware, adopt renewable energy for data center operations, or use cloud computing. These are business decisions, nothing to do with hardcore technology per se, and foster organizational practices like training AI teams on efficiency and incentivizing energy goals.

Department-wise, you can incentivize by setting energy goal targets for the department. To conclude this session, overall, this module and the course—Clean AI—focus on making AI energy-efficient and managing the data flood.This data flood is something that is here to stay and will only increase.A medical scan of a single organ creates 10 GB of raw data each second; every minute, 400 hours of video are uploaded on YouTube; 325,000 new malware files appear every day; 50-plus billion IoT devices collect data; 2.3-plus billion people use smartphones and collect data; Amazon offers 550-plus billion products. You can go on and on, but at the end of the day, it is all about data, data, data—and data moving on the net consumes energy.

Making AI energy-efficient and managing the data flood is the goal or purpose of what we can call Clean AI. With that I will like to conclude my overall course of this

foundation of digital transformation. It is a transformation journey that has already started and will continue—we cannot stop technology, as we said, but the message is: do not—think about the job; think about the people, because with technology, jobs will change and vanish, but people will stay. Thank you very much.