U.S. Department of Energy
Advanced Research Projects Agency – Energy (ARPA-E)

Request for Information (RFI)
DE-FOA-0002495
on
Increasing Data Center Energy Efficiency

Introduction

The purpose of this RFI is to solicit input for a potential future ARPA-E research program focused on novel, potentially transformative technical opportunities, and approaches to **increase the energy efficiency of data centers.**

Online services and online data use have grown exponentially for years and is expected to grow even more so in the upcoming years due to growing trends such as virtual online meetings, artificial intelligence, machine learning, machine vision, and augmented reality[1]. Recent indications also illustrate that we are on the precipice of accelerated machine-to-machine communication also known as Internet of Things (IOT) as cars, aircraft, industrial machines, and household devices will start gathering and communicating data for their operations[2]. These virtual communications occur on an infrastructure supported by millions of servers that are hosted in data centers.

Data centers exist in a variety of sizes and operational models ranging from enterprise, co-location to hyperscale data centers. It is projected that current U.S. data centers consume in excess of 75 billion kWh electricity annually[3] and that innovative technologies such as advances in semiconductors, virtualization and economization have contributed to a relatively modest increase in energy usage compared to the 550% growth in compute instances over the past decade[4]. However, the semiconductor transistor scaling trend which consistently delivered operational energy savings during this time period, or “Moore’s law”, reached an inflection point in 2016, and it is therefore expected that data center energy usage will rise as future data communications demand is expected to grow exponentially[5].

The goal of the potential program is to support transformational technologies that can increase the energy efficiency of current and future data centers over as wide an application range as possible. Examples of potential areas of interest are disruptive innovations in efficient and reliable power supply, data

[1] Renee Obringer, Benjamin Rachunok, Debora Maia-Silva, Maryam Arbabzadeh, Roshanak Nateghi, and Kaveh Madani. "The overlooked environmental footprint of increasing Internet use." Resources, Conservation and Recycling 167: 105389.

[2] “2020 Global Networking Trends Report.” Cisco, Cisco, 4 May 2020, https://www.cisco.com/c/en/us/solutions/enterprise-networks/networking-technology-trends.html

[3] Arman Shehabi, Sarah Smith, Dale Sartor, Richard Brown, Magnus Herrlin, Jonathan Koomey, Eric Masanet, Nathaniel Horner, Inês Azevedo, and William Lintner. "United States Data Center Energy Usage Report." (2016).

[4] Masanet, Eric, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. "Recalibrating Global Data Center Energy-Use Estimates." Science 367, no. 6481 (2020): 984-986.

[5] Fulton, Scott. “After Moore's Law: How Will We Know How Much Faster Computers Can Go?” Data Center Knowledge, 21 Dec. 2020, www.datacenterknowledge.com/supercomputers/after-moore-s-law-how-will-we-know-how-much-faster-computers-can-go.

processing, thermal management, server, rack, or building designs[6]. These technologies could be developed as stand-alone component innovations or be considered as part of system-level approaches that span multiple component innovations as a system.

Traditionally, data center cooling has been one of the significant contributors to data center inefficiencies and its performance significantly climate dependent[7,8]. Trends in the data center space such as higher power processors, increased power density per rack and the emergence of Edge data centers, which are placed closer to the customer to ensure low latency connections and high security for critical applications, are therefore expected to further exacerbate cooling and energy challenges[9,10]. Advanced cooling methods such as two-phase evaporation, and bio-inspired transpiration cooling[11] have been explored for some time through implementations that include immersion cooling[12] or specialized cold plate configurations[13] but have only found limited commercial adoption. Cost and real or perceived reliability concerns have been mentioned as barriers for implementation of some of these new technologies in data centers.

Broad applicability can be improved by exploring technologies that deliver high performance at low cost and can operate efficiently, independently of the climate and/or water availability of the data center size or location. Technologies could explore the use of energy storage (thermal or electric) to take advantage of fluctuating demand/response and/or diurnal cycles.

Thermal management or cooling of a data center can arbitrarily be separated in three different optimization challenges and/or in a single system optimization challenge:

i. Move heat from current and future server chipsets to coolant efficiently
ii. Move heated coolant from server to the heat rejection system with high reliability
iii. Data center heat rejection systems that can reject heat efficiently or re-use the heat economically for auxiliary services that augment or benefit the data center

An important objective of the data center operator is to have high availability, which is defined as the percentage of time the data center is available[14]. When the data center is considered as a system, failure mode and effect analysis can be used to assess the likelihood of occurrence, likelihood of detection, and

[6] Aviv, Dorit, and Forrest Meggers. "Cooling oculus for desert climate–dynamic structure for evaporative downdraft and night sky cooling." Energy Procedia 122 (2017): 1123-1128.

[7] Barroso, Luiz André, Urs Hölzle, and Parthasarathy Ranganathan. "The datacenter as a computer: Designing warehouse-scale machines." Synthesis Lectures on Computer Architecture 13, no. 3 (2018): i-189.

[8] Alkharabsheh, Sami, John Fernandes, Betsegaw Gebrehiwot, Dereje Agonafer, Kanad Ghose, Alfonso Ortega, Yogendra Joshi, and Bahgat Sammakia. "A brief overview of recent developments in thermal management in data centers." Journal of Electronic Packaging 137, no. 4 (2015).

[9] Gordón, Andrea, and Gustavo Salazar-Chacón. "DRP Analysis: Service Outage in Data Center due to Power Failures." In 2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 0182-0187. IEEE, 2020.

[10] Fleischer, A.S., 2020. Cooling our insatiable demand for data. *Science*, *370*(6518), pp.783-784

[11] Huang, Gan, et al. "Experimental investigation of self-pumping internal transpiration cooling." International Journal of Heat and Mass Transfer 123 (2018): 514-522.

[12] Shah, Jimil M., Richard Eiland, Ashwin Siddarth, and Dereje Agonafer. "Effects of mineral oil immersion cooling on IT equipment reliability and reliability enhancements to data center operations." In 2016 15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), pp. 316-325.

[13] Ramakrishnan, Bharath, Cong Hiep Hoang, Sadegh Khalili, Yaser Hadad, Srikanth Rangarajan, Arvind Pattamatta, and Bahgat Sammakia. "Experimental Characterization of Two-Phase Cold Plates Intended for High Density Data Center Servers Using a Dielectric Fluid." Journal of Electronic Packaging (2021).

[14] Ascierto, R., Lawrence, A., "Uptime Institute Global data center survey 2020", UII-38 v1.1, July 2020

severity of impact of potential failure modes of components and subsystems[15]. Generally, the lowest of the individual subsystem ratings determines the overall tier classification of the datacenter. ARPA-E would be interested to learn if new proposed transformational technologies would require health monitoring sensors and/or mechanisms for fail-safe operation, or advanced controls co-designed to ensure they have the potential to enable data center availability comparable or better than achieved with current technologies. In addition, it would be interesting to learn if prognostic health management and condition-based maintenance could reduce server redundancy and further improve energy efficiency.

Another potential approach to increase total energy efficiency of data centers is to re-use the waste heat. Waste Heat Utilization for data centers has been an increasingly important area and increasingly encouraged and, in some cases, being explored to be mandated by local governments[16]. Although connecting a data center to a district heating system or green houses through co-location has been demonstrated, the mismatches in temperature, heat exchanger inefficiencies and operational offsets in energy demand and supply can be challenging[17,18].

As a potential program seeks heat rejection efficiency to be improved, heat could come out of the data center at higher temperatures than before (potentially slightly below typical chipset operating temperatures (70-90°C)). ARPA-E is interested to learn if energy at these temperatures (higher exergy/quality than current state-of-the-art) could potentially lead to new insights for energy re-use that would not need to rely on co-location of other heat demanding industries.

ARPA-E would be interested to learn if transformational technologies can reach additional performance if they were designed and/or co-optimized together. Such an increase could lead to an inflection point where the performance/cost ratio of new technologies would exceed baseline server technology at the data center level, and hence would be tech2market viable if simultaneously high reliability could be achieved.

The envisioned outcome of a potential program would deliver maturation of transformational technology concepts that display high potential performance and broad tech2market viability such that a leap in data center energy efficiency can be realized.

[15] Dai, Jun, Diganta Das, Michael Ohadi, and Michael Pecht. "Reliability risk mitigation of free air cooling through prognostics and health management." Applied Energy 111 (2013): 104-112.

[16] Holla, K. "Waste Heat Utilization is the Data Center Industry's Next Step Toward Net-Zero Energy", datacenterfrontier.com, August 2021, https://datacenterfrontier.com/waste-heat-utilization-data-center-industry

[17] Amalfi, Raffaele L., Filippo Cataldo, and John R. Thome. "An optimization algorithm to design compact plate heat exchangers for waste heat recovery applications in high power datacenter racks." In 2019 18th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), pp. 159-166.

[18] Wahlroos, Mikko, Matti Pärssinen, Jukka Manner, and Sanna Syri. "Utilizing data center waste heat in district heating–Impacts on energy efficiency and prospects for low-temperature district heating networks." Energy 140 (2017): 1228-1238.

Please carefully review the REQUEST FOR INFORMATION GUIDELINES below. Please note, in particular, that the information you provide will be used by ARPA-E solely for program planning, without attribution. **THIS IS A REQUEST FOR INFORMATION ONLY. THIS NOTICE DOES NOT CONSTITUTE A FUNDING OPPORTUNITY ANNOUNCEMENT (FOA). NO FOA EXISTS AT THIS TIME.**

Purpose and Need for Information

The purpose of this RFI is solely to solicit input for ARPA-E consideration to inform the possible formulation of future research programs. ARPA-E will not provide funding or compensation for any information submitted in response to this RFI, and ARPA-E may use information submitted to this RFI without any attribution to the source. This RFI provides the broad research community with an opportunity to contribute views and opinions.

REQUEST FOR INFORMATION GUIDELINES

No material submitted for review will be returned and there will be no formal or informal debriefing concerning the review of any submitted material. ARPA-E may contact respondents to request clarification or seek additional information relevant to this RFI. All responses provided will be considered, but ARPA-E will not respond to individual submissions or publish publicly a compendium of responses. **Respondents shall not include any information in the response to this RFI that could be considered proprietary or confidential.**

Depending on the responses to this RFI, ARPA-E may consider the rapid initiation of one or more funded collaborative projects to accelerate along the path towards commercial deployment of the energy technologies described generally above.

Responses to this RFI should be submitted in PDF format to the email address **ARPA-E-RFI@hq.doe.gov** by **5:00 PM Eastern Time on 4/30/2021**. Emails should conform to the following guidelines:

- Please insert “Response to RFI on Data Center Energy Efficiency - <your organization name>” in the subject line of your email
- In the body of your email, include your name, title, organization, type of organization (e.g. university, non-governmental organization, small business, large business, federally funded research and development center (FFRDC), government-owned/government-operated (GOGO), etc.), email address, telephone number, and area of expertise.
- Responses to this RFI are limited to no more than 10 pages in length (12-point font size).
- Responders are strongly encouraged to include preliminary results, data, and figures that describe their potential innovations.

Targeted Questions Seeking Specific Feedback

Please provide responses and information about any of the following. ARPA-E does not expect any one respondent to answer all, or even many, of these prompts. Simply indicate the question number in your response. Citations are encouraged as appropriate. When possible please be as quantitative as possible, in particular with potential energy savings and carbon footprint reduction, economics and technological advancements, and other areas of direct significance to ARPAE mission. Respondents are also welcome to address other relevant avenues/technologies that are not outlined below.

A. Impact and technical metrics

1. There are many data centers in the US and the sector is expected to continue to grow rapidly. Which data center types should ARPA-E focus on to have the most impact on energy consumption reduction?
2. What are the implications of the Edge data center trend? Is this trend significant and are there other key trends in data center design, rack power density and/or energy efficiency that we should be aware of?
3. Could small modular data centers be as efficient as large data centers? And if so, what would be required for this to be realized?
4. If a technology program would be launched, what would be technical metrics to consider, and why? And what would be considered transformational targets? ARPA-E could consider several potential metrics to drive impactful R&D in this area. One is PUE x ITUE (TUE). Another is data center computational efficiency – $\text{Gflops/watt}_{\text{data-center}}$. What are representative metrics for state-of-the-art technologies in context of each? Please include any references that would be applicable. Are there alternative/better metrics that ARPA-E should consider? Why/why not?

B. Technical approaches to improving data center energy efficiency

1. What are currently, and projected to be, the largest drivers for inefficiencies in current and future data centers (i.e. efficiencies of power supply, data processing, data center cooling, etc.)?
2. What are effective ways to benchmark the computational performance capability of a data center such that this can be normalized by its power usage (i.e. $\text{Gflops/watt}_{\text{data-center}}$)?
3. Data Center cooling can involve multiple steps, i.e. getting heat of the chip, heat transport and/or heat rejection or re-use. Which of these areas do you identify as having the largest potential for innovation and can be these be innovated as component technologies or do they need to be co-designed as a system? Are there new insights, e.g. use of additive modalities such as additive ceramics ($Al_2O_{3,}$ AlN, etc.), aluminum or copper that could assist in concepts that allows coolants to connect to the chipset more efficiently? What would be the entitlement?
4. Cost and reliability have been mentioned as barriers for implementation of new efficient technologies in data centers. To overcome these barriers and become as economically viable as baseline systems, should a performance increase be targeted such that an improved performance / cost ratio is achieved? If so, approximately, how much additional performance (GFlops, power/rack or other) needs to be achieved for every dollar technology cost added?
5. Operational availability is a key metric for data centers. Are there new transformational technologies concepts in the areas of health monitoring sensors and/or fail-safe mechanisms or controls that could contribute to achieving system operational availability comparable or better than state of the art?
6. In order to have a transformational impact, what would it take for waste heat from data centers to be used economically and practically without reliance on co-location of other industries? At what temperature would the exit air/coolant become a valuable resource at scale? Are there opportunities for the waste heat to be used to further reduce the energy footprint of other systems in the data center?
7. Are there greater performance gains to be had from component-level or system-level technology approaches? What are specific system-level gains that couldn't be achieved with siloed component development efforts?
8. Do you currently use system-level economic, reliability or energy-optimization design tools for your data center, and if so, what are their limitations? Should ARPA-E consider additional resources to development/further augmentation of such software/hardware tools?
9. Are there any other promising technical approaches/areas not included in this document that ARPA-E should consider for improving energy efficiency of datacenters?