Efficient Computing

Comparison of Cryogenic Versus Non-Cryogenic Computing Systems by Cooling Power

Motivation

We are interested in comparing computing systems based on semiconductor versus superconductor materials. Computing systems based on superconductor materials are likely to be hybrid systems including cryogenic semiconductor materials—like “cryo CMOS” or “cold CMOS,” for example—so it may be more accurate to say we are interested in comparing cryogenic and non-cryogenic computing systems. What would motivate a buyer—such as a data center operator—to purchase a cryogenic system over an incumbent non-cryogenic system? Putting aside factors like software compatibility and optimization, which are more mature for non-cryogenic systems, this memo will consider system power efficiency.

Within this scope, a logical buyer would purchase the system that draws less power. In a data center scenario, electricity for IT and cooling equipment is a significant operating expense, so the more efficient system is likely to be more economical subject to capital costs and the life of the center. From an environmental perspective, the more efficient system is likely to have less of a negative impact on the overall environment by reducing greenhouse gas emissions attributable to electricity generation. Cryogenic systems may also have less of a negative impact on the local environment than non-cryogenic systems—for example, by utilizing liquid nitrogen rather than water in a desert.

Calculating Efficiency Breakpoints

Cryocooling generally requires more power than non-cryogenic cooling, so the cryoelectronics will need to be more efficient than the non-cryogenic electronics to compensate for (and ultimately overcome) this inherent disadvantage. The breakeven point is straightforward:

$$Electronics\ Power\ +\ Cooling\ Power\ =\ Cryoelectronics\ Power\ +\ Cryocooling\ Power$$

Which we will write as:

$$P_{electronics} + P_{cooling} = P_{cryoelectronics} + P_{cryocooling} \quad (1)$$

Let us represent *Cryoelectronics Power* as *Electronics Power* divided by some *Efficiency Factor*, *E*, because we are interested in the relative efficiency of the electronics versus cryoelectronics.

$$P_{cryoelectronics} = \frac{P_{electronics}}{E} \quad (2)$$

Thus Equation (1) becomes:

$$P_{electronics} + P_{cooling} = \frac{P_{electronics}}{E} + P_{cryocooling} \quad (3)$$

The power required to cool the systems can be determined using heat dissipation and the power required for a given cooling capacity at a given temperature (W/W @ T). Let *N* and *C* be these values for the non-cryogenic and cryogenic systems, respectively.

$$P_{cooling} = NP_{electronics} \quad (4)$$

$$P_{cryocooling} = CP_{cryoelectronics} = C\frac{P_{electronics}}{E} \quad (5)$$

Substituting (4) and (5) into (3) we have:

$$P_{electronics} + NP_{electronics} = \frac{P_{electronics}}{E} + C\frac{P_{electronics}}{E}$$

Which simplifies to:

$$(1 + N)P_{electronics} = (1 + C)\frac{P_{electronics}}{E}$$

$$(1 + N)P_{electronics}E = (1 + C)P_{electronics}$$

$$E = \frac{(1+C)P_{electronics}}{(1+N)P_{electronics}}$$

$$E = \frac{1+C}{1+N} \quad (6)$$

So, intuitively, when comparing computing systems at different temperatures, we can determine the breakeven efficiency factor by comparing the cooling power required to operate those systems at their respective temperatures.

Cryocooling Power

To determine cryocooling power, we will begin with the ideal (Carnot) minimum power required to cool the cryogenic system from room temperature to operating temperature (Table 1). Then we will compare that value to the power requirements of commercially available cryocoolers (Table 2). We could simply reference real-world W/W values from commercial spec sheets, but this way will show us how efficient commercial cryocoolers are. The result is a range of W/W values we can use as parameters for more conservative or more optimistic estimates.

Table 1: Ideal Minimum Cryocooler Power

Ideal (Carnot) Minimum Cryocooler Power	
Ambient Temperature (°C)	20
Ambient Temperature (K)	293.15
Target Temperature (K)	77
Cooling Capacity (W)	1
Minimum Power Requirement (W)	2.81

Table 2: Real Cryocooler Power

Real Cryocooler Power					
Cryocooler	Lift @ 77 K (W)	Input Power (W)	Ideal Power Requirement (W)	Efficiency	W/W
CryoTel DS MINI	1.8	45	5.05	11.23%	25.00
CryoTel MT	5	80	14.04	17.54%	16.00
CryoTel CT	11	160	30.88	19.30%	14.55
CryoTel GT	16	240	44.91	18.71%	15.00
CryoTel DS30	32	480	89.83	18.71%	15.00
Stirling SPC-1 Cryogenerator	1000	10700	2807.14	26.23%	10.70

Cooling Power

We will use data from industrial server coolers as well as consumer coolers for high-end PCs—typically used for gaming or content creation—to estimate server cooling requirements (Table 3). We include consumer data because it is readily available. One argument why server cooling may be better than PC cooling is that industrial products tend to be more sophisticated or optimized than consumer products. One argument why PC cooling may be better than server cooling is that consumers may have relaxed size constraints—coolers in PC cases may be able to take up absolutely or relatively more space than coolers in densely packed server racks, for example—and larger fans, greater radiator surface area, etc. often result in more cooling.

Note to Reader: The results of this analysis are not nearly as sensitive to non-cryogenic cooling power as they are to other parameters, so a variance of a few Watts for the non-cryogenic cooler will not significantly skew the conclusions. Also, we assume the ambient environment is identical across cryogenic and non-cryogenic scenarios. In other words, we will not consider facility air conditioning to be part of cooling estimates.

Table 3: Real Cooler Power

Real Cooler Power				
Component Pair	Market	Power (W)	W/W	Source/Notes
Boyd 4U CDU Cooling Capacity Boyd 4U CDU Power Consumption	Industrial	80000 800	0.0100	Boyd 4U Coolant Distribution Unit specifications.
AMD Ryzen 9 3950X OC Deepcool AK620 (2 x FK120 Fans)	Consumer	198 2.88	0.0145	Gamers Nexus experiment. Fan specifications.
NVIDIA GeForce RTX 4090 FE 2 x Nidec AD4A31K04/AD4A31K05 Fans	Consumer	450 15.12	0.0336	Assume fans draw max rated power at max rated TDP.
Dynatron L32 Cooling Capacity Dynatron L32 Max Power Draw	Consumer	455 15.6	0.0343	AIO liquid cooler for AMD EPYC 9004 Series server processors.
ASUS RS720A-E12-RS24U Max Power Supply 4 x 84 W Fans	Industrial	2600 336	0.1292	Server specifications. Fan specifications (12 V) (7 A) visible at 5:46.

Again, the result is a range of W/W values we can use as parameters for more conservative or more optimistic estimates. The significantly higher W/W value for the industrial air cooler may indicate an outlier, but this is immaterial because we use the best case for non-cryogenic cooling for calculations after Table 4.

Real Efficiency Breakpoints

We can plug the W/W values from Tables 2 & 3 into Equation (6) to determine real efficiency breakpoints (Table 4).

Table 4: Efficiency Breakpoints over a Range of Real Coolers and Cryocoolers

	Cryo		
Non-Cryo	Worst (~11.2%)	Median (~18.7%)	Best (~26.2%)
Worst	23.0	14.2	10.4
Median	25.2	15.5	11.3
Best	25.7	15.8	11.6

Breakpoints appear to be dominated by cryocooler efficiency (grouping occurs around cryocooler parameters rather than cooler parameters). So, depending on the cryocooler, cryoelectronics at 77 K need to be 10-26 times more efficient than ~room-temperature electronics to compensate for additional cooling.

Varying Efficiency Factor

Rather than calculate the breakeven efficiency, it may be helpful to illustrate the effect efficiency factor has on system power. Tables 5 & 6 model the most conservative scenario (best-case non-cryo, worst-case cryo).

Table 5: Parameters for Varying Efficiency Factor

Parameters	
Electronics Power (W)	1000
Cooling Power (W/W)	0.0100
Ideal Minimum Cryocooling Power (W/W)	2.81
Real Cryocooling Efficiency	11.23%
Real Cryocooling Power (W/W)	25

Table 6: Varying Efficiency Factor

Varying Efficiency Factor								
E	Electronics (W)	Cooling (W)	Total Non-Cryo (W)	Cryoelectronics (W)	Cryocooling (W)	Total Cryo (W)	Cryo/Non-Cryo	Non-Cryo/Cryo
1	1000	10	1010	1000	25000	26000	25.74	0.04
10	1000	10	1010	100	2500	2600	2.57	0.39
20	1000	10	1010	50	1250	1300	1.29	0.78
25	1000	10	1010	40	1000	1040	1.03	0.97
26	1000	10	1010	38	962	1000	0.99	1.01
40	1000	10	1010	25	625	650	0.64	1.55
60	1000	10	1010	17	417	433	0.43	2.33
80	1000	10	1010	13	313	325	0.32	3.11
100	1000	10	1010	10	250	260	0.26	3.88
200	1000	10	1010	5	125	130	0.13	7.77
300	1000	10	1010	3	83	87	0.09	11.65
400	1000	10	1010	3	63	65	0.06	15.54
500	1000	10	1010	2	50	52	0.05	19.42
1000	1000	10	1010	1	25	26	0.03	38.85
5000	1000	10	1010	0	5	5	0.01	194.23
10000	1000	10	1010	0	3	3	0.00	388.46
50000	1000	10	1010	0	1	1	0.00	1942.31
100000	1000	10	1010	0	0	0	0.00	3884.62

So, for example, in the most conservative scenario (best-case non-cryo, worst-case cryo), the cryogenic system would be ~4 times more efficient than the non-cryogenic system if the cryoelectronics operated 100 times more efficiently than the ~room-temperature electronics.

The breakeven point according to Equation (6) is visible in the transition from red to green as well as in the first row, where *E* = 1:

$$E = \frac{1+C}{1+N}$$

$$E = \frac{1+25}{1+0.01}$$

$$E \approx 25.74$$

Extrapolations from the Literature

Going back to our original interest in comparing computing systems based on semiconductor versus superconductor materials, let's consider two groups of superconducting materials.

Low-temperature superconductors (LTS), often used to make qubits and control circuits for quantum computing, are typically cooled by liquid helium to operate at or below 4 Kelvin. Niobium is a commonly used LTS.

High-temperature superconductors (HTS), often used in clean energy applications like high-capacity cables for smart grids and the world's most powerful magnet systems, are typically cooled by liquid nitrogen to operate at or above 77 Kelvin, but include materials that transition at much higher temperatures, up to room temperature. YBCO is a commonly used HTS and a member of the cuprate family of superconductors.

Note to Reader / Bias: You may notice that the calculations in this memo have been based on HTS. Ambature's patented a-axis HTS materials enable semiconductor foundries to fabricate HTS circuits at scale using standard semiconductor equipment—etching, lithography, sputtering, etc. The unique orientation of a-axis cuprates versus conventional c-axis cuprates allows us to overcome historical barriers of HTS device design and manufacturing, making HTS a viable alternative to LTS in many applications and enabling new applications of superconductors in science, industry, and defense.

Many research papers estimate the efficiency of LTS circuits relative to semiconductor circuits. We can use the difference between LTS and HTS cryocooling power to estimate the efficiency of HTS circuits relative to semiconductor circuits. Estimates should be reasonable because cryocooling power dominates cryogenic system power, especially in LTS scenarios, where cryocooling power dwarfs cryoelectronics power. Tables 7 & 8 compare semiconductor, LTS, and HTS circuits across several applications using papers' specified cooling assumptions. Table 7 uses the worst-case HTS cryocooler efficiency from Table 2 while Table 8 uses the best case.

Table 7: Efficiency of HTS vs LTS Versus Semiconductor Circuits (Conservative)

Efficiency of HTS Versus LTS Versus Semiconductor Circuits (Conservative)										
Paper	Year Published	SC Logic	Semiconductor Baseline	Application	Baseline / LTS	LTS Cooling Overhead (W/W)	Clock Speed	Baseline / HTS	HTS Cooling Overhead (W/W)	HTS / LTS
1	2019	RQL	16nm CMOS	SHA-256 Accelerator	46	300	Not Specified	552	25	12
1	2019	Not Specified	16nm CMOS	64-bit Add	9.86	300	Not Specified	118.32	25	12
1	2019	Not Specified	16nm CMOS	64-bit Multiply	9.54	300	Not Specified	114.48	25	12
1	2019	Not Specified	16nm CMOS	64-bit RF Load	1	300	Not Specified	12	25	12
1	2019	Not Specified	16nm CMOS	Off-Chip Interconnect	30,000	300	Not Specified	360,000	25	12
2	2021	AQFP	2017 7nm @ VDD = 0.8 V	Inverter, Various	79.21	1000	100 kHz - 5 GHz	3,049	25	40
3	2015	RSFQ	CMOS FPGA @80MHz	Collatz Conjecture	7,300	1000	37 GHz	292,000	25	40
3	2015	AQFP	CMOS FPGA @80MHz	Collatz Conjecture	7,300,000	1000	5 GHz	292,000,000	25	40

So, for example, in the most conservative scenario, we would expect an HTS implementation of a SHA-256 accelerator to be 12 times more energy-efficient than an LTS implementation of a SHA-256 accelerator, and 552 times more energy-efficient than a 16nm CMOS implementation of a SHA-256 accelerator.

Table 8: Efficiency of HTS vs LTS Versus Semiconductor Circuits (Optimistic)

Efficiency of HTS Versus LTS Versus Semiconductor Circuits (Optimistic / Large Scale)										
Paper	Year Published	SC Logic	Semiconductor Baseline	Application	Baseline / LTS	LTS Cooling Overhead (W/W)	Clock Speed	Baseline / HTS	HTS Cooling Overhead (W/W)	HTS / LTS
1	2019	RQL	16nm CMOS	SHA-256 Accelerator	46	300	Not Specified	1,290	11	28
1	2019	Not Specified	16nm CMOS	64-bit Add	9.86	300	Not Specified	276.45	11	28
1	2019	Not Specified	16nm CMOS	64-bit Multiply	9.54	300	Not Specified	267.48	11	28
1	2019	Not Specified	16nm CMOS	64-bit RF Load	1	300	Not Specified	28	11	28
1	2019	Not Specified	16nm CMOS	Off-Chip Interconnect	30,000	300	Not Specified	841,121	11	28
2	2021	AQFP	2017 7nm @ VDD = 0.8 V	Inverter, Various	79.29	1000	100 kHz - 5 GHz	6,777	11	93
3	2015	RSFQ	CMOS FPGA @80MHz	Collatz Conjecture	7,300	1000	37 GHz	682,243	11	93
3	2015	AQFP	CMOS FPGA @80MHz	Collatz Conjecture	7,300,000	1000	5 GHz	682,242,991	11	93

Given the breakeven points from Table 4 fall between 10-26, these estimates in the hundreds, thousands, and millions (*including* cooling overhead) indicate that HTS and hybrid semiconductor/superconductor cryoelectronics warrant further development to create far more efficient systems than ~room-temperature and LTS incumbents, particularly when researchers illustrate how compute needs are "exploding" (Figure 1) and new energy-efficient options are needed to prevent compute energy from eclipsing global energy production by 2040 (Figure 2).

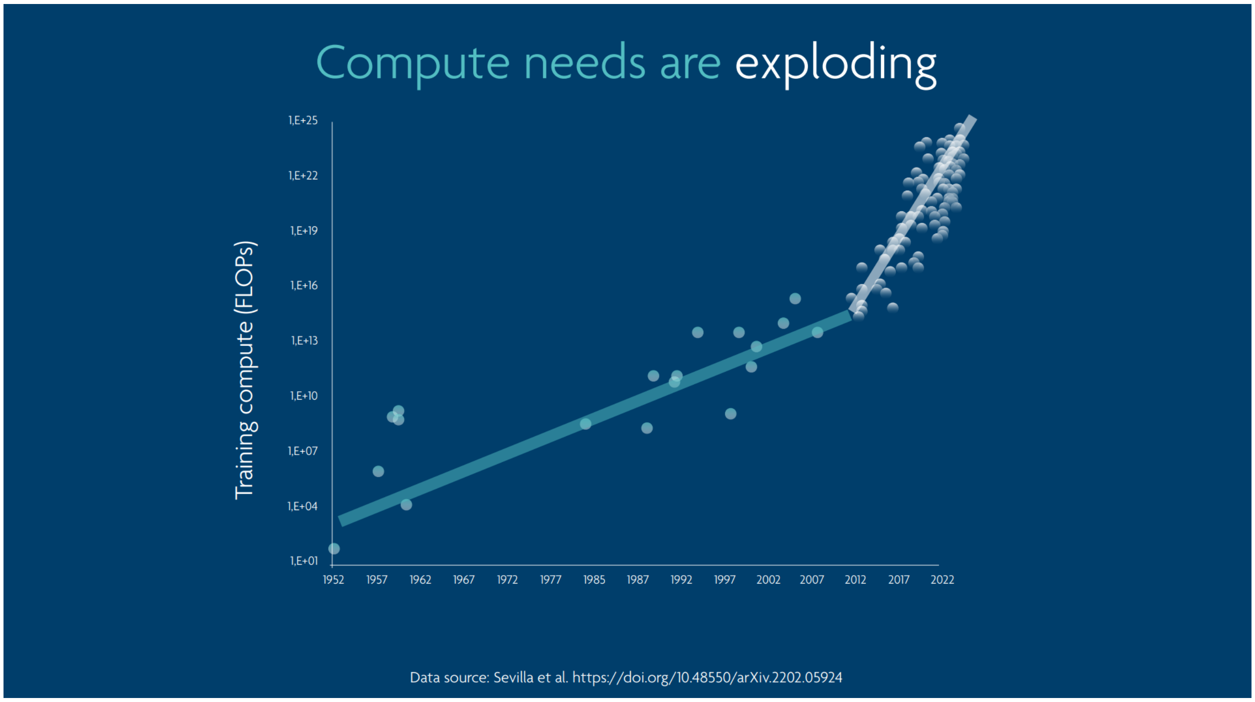

Figure 1: Compute Needs Are Exploding (Sevilla et al. via imec)

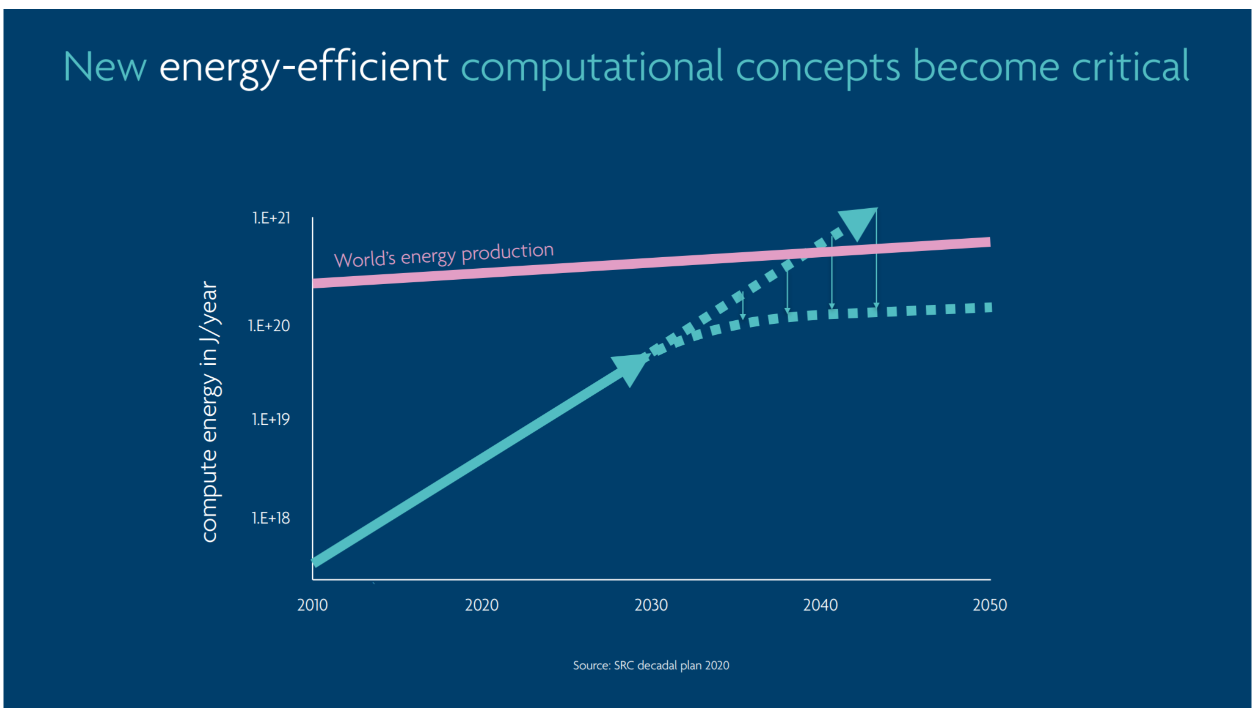

Figure 2: Compute Energy & Global Energy Production (SRC via imec)

Superconducting Supercomputers

The most efficient cryocooler in Table 2 is also the most suitable for large-scale applications given it has the highest cooling capacity. A supercomputer would be a good example of a large-scale application. In 2013, Holmes, Ripple, and Manheimer estimated that exascale computing with superconducting Reciprocal Quantum Logic (RQL) would require 867 W of cryoelectronics (Table 9) in a ~2 MW system (Table 10).

Table 9: Exascale Supercomputer - Cryoelectronics

Exascale Supercomputer - Cryoelectronics		
	Logic	Total *
RQL W/EFLOP/s (kW)	0.260	0.867
LTS Cryocooling (W/W)	395	395
LTS RQL W/EFLOP/s (kW)	103	343
HTS Cryocooling (W/W)	11	11
HTS RQL W/EFLOP/s (kW)	3	10
% Improvement		97.05%
Factor Improvement		34

* Logic, Memory, Interconnects, Heat Leaks

Table 10: Exascale Supercomputer - System Power

Exascale Supercomputer - System Power								
	Conservative		Moderately Conservative		Moderately Optimistic		Optimistic	
	LTS	HTS	LTS	HTS	LTS	HTS	LTS	HTS
Refrigeration (kW)	343	10	343	10	343	10	343	10
Non-Cryo Improvement Factor **	N/A		2.5		6		7.5	
Other (kW)	1600	1600	640	640	267	267	213	213
Total (kW)	1943	1610	986	650	616	277	564	223
% Improvement	17.14%		34.04%		55.05%		60.38%	
Factor Improvement	1.21		1.52		2.22		2.52	
GFLOPS/W	500	603	986	1494	1578	3510	1723	4348

** Estimated improvement in non-cryogenic component efficiency between ~2013 and ~2023:

(2.5) For example, the NVIDIA GeForce RTX 4090 is roughly 5 times more performant than the NVIDIA GeForce GTX Titan and draws roughly double the power.

(6) For example, the AMD Ryzen 9 7950X3D is roughly 6 times more performant than the Intel Core i7-4960X and draws less power.

(7.5) For example, in 2008 an exascale computer was estimated to require more than 600 MW. Using projected 2015 technologies, the estimate was still 150 MW. The DoE goal of 1 EFLOP/s for 20 MW equates to 50 GFLOP/J. Frontier (1.194 EFLOPS) achieved 52.59 GFLOPS/W in June 2023.

(30) For example, the most efficient supercomputer in June 2013 achieved 2 GFLOP/J, while the most efficient supercomputer in June 2023 (Henri) achieved 65.40 GFLOPS/W (2.88 PFLOPS), followed by the Frontier Test & Development System (TDS) at 62.20 GFLOPS/W, a test rack of Frontier.

Holmes et al. estimated the 2 MW system would achieve 500 GFLOP/J (GFLOPS/W), 10 times the US Department of Energy's (DoE) exascale computing goal of 50 GFLOP/J. As of June 2023, Frontier is the No. 1 supercomputer in the TOP500 list at 1.194 EFLOPS and No. 6 in the Green500 list at 52.59 GFLOPS/W, meaning the RQL system posited in 2013 is estimated to be 10 times more efficient than the world's most efficient supercomputer today. Implementing a supercomputer with alternatives to RQL—like Adiabatic Quantum-Flux Parametron (AQFP) logic, for example—could be even more efficient.

You may notice that power consumption in this system is dominated by "Other" system components, including "room temperature interconnect drivers and receivers, power supplies, and storage memory." In an attempt to modernize these estimates, decadal improvement factors are estimated in Table 10 based on approximate CPU, GPU, and supercomputer improvements over the last 10 years. If these factors are reasonable, GFLOPS/W estimates are significantly improved and HTS implementations offer greater benefits over LTS ones.

Further Considerations

The cost of energy is not the only meaningful consideration when comparing LTS and HTS systems. Liquid nitrogen is more abundant and less expensive than liquid helium, for example (10-20 times in our own lab experience). The scarcity of accessible helium may become a barrier to widespread adoption of LTS technologies. By contrast, accessible nitrogen is effectively limitless, and liquid nitrogen is used regularly by industry.

Complementary superconducting technologies can lead to additional efficiencies in supercomputer or data center scenarios. For example, liquid nitrogen-cooled busbars made of HTS tapes can carry supercurrents throughout the center for a 10-15% reduction in power losses. The liquid nitrogen can also serve as localized fire protection for IT infrastructure. Superconducting energy storage systems like frictionless flywheel or magnetic systems can balance loads and provide emergency power to the center.

We generally discuss superconducting computing within the scope of large-scale industrial applications like supercomputers and data centers. While we expect this to hold true in the medium term, long-term opportunities in the consumer space may exist given current market trends. High-end PCs can cost thousands of dollars, employ liquid cooling, and draw upwards of 700 W (excluding transient power spikes) depending on consumer CPU/GPU combinations. As superconducting circuit technology matures alongside cold semiconductor technology (e.g. for quantum computing control circuits) and small, efficient, closed-cycle cryocoolers like those used for space applications, a shift from higher-power semiconductor electronics with lower-power cooling solutions to lower-power hybrid cryoelectronics with higher-power cooling solutions is not outside the realm of possibility. Software compatibility/support from companies like Intel, AMD, NVIDIA, etc. would likely play a larger role than overall system efficiency in determining whether superconductors are adopted into consumer markets.