# 通过创建学术云计算基础设施，挽救数百万美元投向规划不当的高性能计算系统的尝试：一段充满错误与迟来教训的历程。

Marek Michalewicz[1]

摘要— 2015年，华沙大学数学与计算建模跨学科中心（ICM）建成了一座现代化数据中心，并作为一项总额为1.68亿波兰兹罗提（3600万欧元）的OCEAN项目的一部分，部署了三套大型高性能计算系统。其中一些系统设计不当、架构糟糕，在五年的使用寿命中几乎没有带来投资回报。本文介绍了一项为期两年的深入工作，旨在将其中两套高性能计算系统重新设计为一种名为A-CHOICeM（Akademicka CHmura Obliczeniowa ICM[2]）的混合多云解决方案。

此举旨在将ICM典型高性能计算系统的用户群，从约200至500人扩大到华沙地区所有高等教育机构中约10万名潜在的普通学术用户。

该解决方案的主要特点包括：将本地部署的ICM云与多个公有云提供商集成，构建针对特定学术用户组量身定制的解决方案，实现容器化，并集成人工智能和量子计算等特殊计算范式。文中介绍了该解决方案的完整设计流程、与供应商开展的竞争性对话，以及该解决方案的最终完整规格。此外，文中还介绍了实施过程中遇到的若干障碍、陷阱和困难，包括“老派”高性能计算管理员的保守态度、大学官僚体制、国家资助政策等。

关键词— 学术云服务、高性能计算云、本地部署的私有学术云、私有云—公有云。

## 一、Introduction

A2015年年底，数学与计算建模跨学科中心（ICM）华沙大学建成了一座现代化的10,000平方米技术空间数据中心（图1），并作为一项总额为1.68亿波兰兹罗提（3,600万欧元）的OCEAN项目的一部分，安装了三套大型高性能计算系统。这些系统包括：Cray XC40（Okeanos）超级计算机；华为集群（Enigma），专门用于借助 Apache Spark 和 Hadoop 进行大数据存储与分析。Enigma由约350台双路插槽、配备12个核心中央处理器的服务器组成，每台服务器配有24 TB本地磁盘空间，总内存为43 TB，磁盘存储为8 PB（图2）。另一套同样来自华为的集群“Topola”用于承载高性能计算负载。该集群由240台服务器组成，共有6,720个核心、23 TB内存，标称计算性能接近300每秒万亿次浮点运算。

尽管在运营的前三年（2016—2018年），Cray Okeanos 和华为 Topola 计算机都得到了充分利用，但 Enigma 华为集群的计算能力利用率从未超过约 20%，存储容量利用率也从未超过约 12%。对于这位于 2018年 出任 ICM 主任的作者而言，这种投资浪费以及大量计算资源和存储资源闲置的状况难以容忍和接受。然而，任何补救措施都需要找到项目自有资金来源，而这类资金向来十分稀缺。2020年 初，ICM 开始创造自有收入，终于出现了机会，也有了充足资金启动升级项目，将 Enigma 和其他一些计算机系统连同大量存储资源改造成私有云服务。



图1 华沙新建的ICM数据中心




图2 ICM数据中心的华为Enigma BigData系统（左）和Okeanos Cray XC40超级计算机。

[1] Marek Michalewicz 于2018年7月至31st 2021年12月担任华沙大学数学与计算建模跨学科中心主任。目前，他是位于波兰华沙的Siranga Sp. z o.o. 创始人兼首席执行官，m.michalewicz@siranga.eu

[2] ICM工程学术计算云

## II. Cloud Services 于 Academia 和 Research

与本地计算资源相比，云计算具有许多显著优势 [1-3]。在高性能计算领域，分布式资源共享最初通过网格计算范式得以实现。美国的 TeraGrid 计划是一个尤为突出的例子 [4,5], ，随后该计划发展为 XSEDE 计划及其基础设施 [6,7]。美国国家科学基金会的这些计划目前由高级网络基础设施协调生态系统（ACCES）项目延续 [8]。网格计算范式在某种意义上具有云计算先驱的特征，曾广泛应用于高性能计算领域和学术计算中心；例如，在波兰，该范式通过 PL-Grid 项目得以实现 [9]。Nimrod/G 是一项解决方案的实例，可在全球范围的计算网格中实现资源管理与共享 [10]。开放数据、开放科学以及共享计算资源和计算基础设施的需求相互交汇，促成了若干将开放数据与云计算范式结合起来的国家级、区域级和国际级计划 [11-18]

高性能计算云计算的概念、技术和解决方案由 UberCloud 的重要倡议[19,20] 以及商业产品提供，这些产品以门户、用户体验解决方案和资源的形式，集成并配置相关解决方案。Rescale [21] 是首个此类商业解决方案，随后还有 XTREME-D [22,23] 和 Ronin [24], 等。一个大规模实验性云计算的重要例子是 Chameleon [25,26] ，由阿贡国家实验室开发。

## III. Problems 的 Purpose 和 Design

Enigma 的利用率极低，从未超过计算资源的 20%，磁盘存储总量的利用率也不超过 12%。这源于多项严重的设计和采购失误：i) 2014 年规划基础设施需求时，曾预期 MapReduce Hadoop 将得到广泛采用，但这一预期从未实现；ii) 由 10 个机架、5 个模块组成的大数据 Enigma 系统在架构设计上存在若干关键缺陷，若不对架构进行重大重新设计，几乎无法根除这些问题：首先，该系统既不支持多用户，也不支持多租户，因为同一模块内的数据（两个机架或约 70 台服务器）可能被同一模块上的其他用户覆盖；此外，通往外部世界的外部数据链路仅为 1 Gbps，这意味着要将用于 Enigma 大数据处理的 8PB 本地磁盘存储填满，需要连续两年多不断导入数据。InfiniBand 互连网络的拓扑结构相当笨拙，既无法快速传输来自集群外部的数据，也无法在集群内部快速传输数据；以太网网络仅用于管理设计，无法承载任何数据流量。需要指出的是，Enigma 的一个模块已用于 OPENAire 项目，ICM 在其中负责提供硬件资源 [1]。

## IV. Problems 的 Internal Resistance 与 SystemsRoadblocks

作者于 2016年末 以副主任身份加入 ICM 后不久，首次意识到高性能计算资源闲置的问题，以及大量公共资金资助的基础设施被极度浪费的惊人现象。当他与同事开始讨论这一问题时，却遭遇了冷漠和不适。技术人员否认这一问题的存在，决策者似乎既不愿负责，也不关心此事。2018年，作者被任命为 ICM 的代理主任后，终于可以启动一个项目，重新设计 Enigma 的用途和配置，并将部分 20+ PB 的孤立存储整合到碎片化程度较低的存储系统中。然而，这说起来容易做起来难。许多障碍和问题都必须加以解决，其中一些包括：

i) ICM 技术管理人员中“老派”高性能计算管理员的保守态度。我们[3] 可以列举一些他们不愿接受变革的原因：

> [START] – 文化阻力：许多高性能计算管理员习惯于沿用传统做法，可能会抵制变革，尤其是在他们觉得自己的专业能力受到贬低或被取代时。– 控制权：高性能计算管理员可能习惯于完全掌控自己的硬件、软件和数据，因此可能不愿将这些关键要素托付给第三方云服务提供商。– 性能：高性能计算应用需要极高的性能和极低的延迟，而由于虚拟化和网络开销，在云环境中可能难以达到这一要求。高性能计算管理员可能担心迁移到云端会对性能产生负面影响。– 成本：云服务提供商通常会按计算资源和带宽使用量收费，对于会产生大量数据的高性能计算工作负载来说，这些费用很快就会累积起来。– 复杂性：云计算会为高性能计算工作流程增加额外的复杂环节，包括网络、安全和数据管理。高性能计算管理员可能会担心管理这些新组件需要投入额外的时间和精力。 [END[4]]

一些 ICM 管理人员和技术人员也表达了上述反对意见。我们持续展开对话，以达成共识，并争取让技术团队中所有经验丰富、但有时固执的成员都参与进来。

该构想的批评者提出的另一个异议是，这个项目完全可以由我们组织内部自行实现。作者对此观点强烈反对，并指出：

[3] 作者向 CgatGPT 提出的问题是：“为什么‘老派’的高性能计算管理员会抵制云计算?

[4] ChatGPT 撰写了 [START] 与 [END] 之间的文本。我自己也会提出同样的论点。

a. 商业云服务提供商的解决方案和技术领先多年，来自这一领域的实体最适合提供我们所寻求的解决方案；
b. 这类开发项目对 ICM 的研究和创新价值不大；c. ICM 的技术人员已经全力投入日常工作和大量研究项目；
d. ICM 不应充当软件开发机构，而应作为研究机构，技术人员的所有精力都应投入到支持研究上，而不是软件开发项目；

e. 我们的员工可以从事后台的“Open Stack”、虚拟化及类似开发工作，但在计量、用户期望与用户体验、庞大新用户群体的即时需求以及前端管理方面缺乏经验。高性能计算从业者知道，管理员与用户偶尔会出现对立，这种情况未必能转化为“用户友好”的产品。

ii) 大学官僚体制。
官僚不喜欢变革和例外。安稳、不受干扰且按部就班的状态才是常态——一旦遇到独特而特殊的情况，他们就会设置障碍，并试图拖慢或搅黄整个流程[27]——本案例正是如此。从2020年6月到2020年12月底，华沙大学公共采购办公室经理花了六个月时间，进行了繁琐且偶尔流于琐碎的调查，才让ICM得以启动“竞争性对话”。采购内容并非“现成的、开箱即用的”产品，而是一种从未在波兰实施过的原创技术解决方案；解决方案提供商必须组建合作伙伴联盟，将多个技术组件整合为一个解决方案；而且只有经验极其丰富的技术厂商才有能力交付这样的方案——这些情况，大学采购办公室并不容易理解和接受。

iii) 项目资金
波兰加入EuroHPC后，HPC的国家资助政策发生了重大变化。大部分资金通过EuroHPC计划拨付，科学与高等教育部将这些活动的协调工作委托给克拉科夫的 Cyfronet[5] 。ICM也没有资格获得所谓的POIR国家资助计划[29] ，因为华沙和马佐夫舍地区被认定为发达地区，而约93%的POIR资金被拨付给欠发达地区。华沙大学也不愿向ICM提供额外资金，因为ICM于2015年完成的OCEAN项目出现了约1,800万波兰兹罗提的资金缺口，华沙大学将其视为ICM的内部债务。耐人寻味的是，ICM作为OCEAN项目的一部分建成的极其现代且极具价值的基础设施，包括数据中心和全部高性能计算基础设施，都被吸收为华沙大学的资产，而这些资产从未用于抵销ICM的现金债务。

自2018年起，ICM 启动了多个项目，这些项目在随后三年产生了足够的收入，用以启动并资助 A-CHOICeM 项目。

iv) 组件之间的技术不兼容性。
Enigma 大数据集群拥有高达 8 PB 的本地磁盘存储，难以与 Oceanos Cray XC 超级计算机的 12 PB HPC Lustre 存储集成。必须强调的是，作为 Oceanos Cray XC40 系统一部分的 Tetyda DDN HPC 存储系统，其配置超出了 ICM 用户的需求。Tetyda 的原始存储容量为 12 PB，截至2021年，实际使用量从未超过 2 PB。在 A-CHOICeM 项目中，我们探讨了如何将这大量闲置的存储容量用于非高性能计算的学术和研究需求。我们还计划将拥有 240 台服务器的 Topola 纳入内部 A-CHOICeM 云基础设施，因为这台利用率极高的计算机所承载的大多数工作负载都是较小的作业，无需超过几台服务器即可完成；其中很大一部分请求来自 PL-Grid 系统，而该系统无论如何都已亟需进行云式升级。

## V. 决定ECISION 将 大数据ONVERT 和IG 高性能计算ATA 及 HPC 系统 YSTEMS 转变为本地部署-PREMISES 云LOUD 计算OMPUTING 系统YSTEM

A-CHOICeM 项目旨在整合所有未充分利用的 ICM 计算资源和存储资源，并使其能够供新型用户使用。

该计划旨在向所有学术用户免费提供内部云资源。但该计划的一项极为重要的特点，是不将 ICM 的私有云资源与公有云资源隔绝开来。我们坚持创建一个“主门户”，供约 100,000 名学术用户访问，最初面向华沙地区的用户；该地区有众多大学和高等教育机构。该门户将连接 ICM 内部的约 5,000–10,000 台虚拟机和 ~20PB 存储资源，同时也允许用户免费访问所有主要提供商的公有云，例如 AWS、Azure、IBM、Green Lake 等。遗憾的是，一些原本颇具实力的潜在解决方案供应商与单一公有云提供商绑定过于紧密，无法提供适用于所有公有云提供商的开放式解决方案。

## VI. 竞OMPETITIVE 争IALOGUE 性ROCESS对话

ICM 遭到了大学公共采购办公室的抵制。从本质上说，那里的官僚人员并未准备好接受这样一个复杂且新颖的项目。他们坚持采用一种更简单但并不充分的招标流程。
申请采用竞争性对话程序的正式理由如下：

“*contracting authority*确认，将根据第*60a*条第*1*款开展竞争性对话程序

[5] 目前已重组并更名为教育与科学部

，以满足《公共采购法》（*PPL*）第*55*条第*1*款第*6*至*9*项规定的条件：

*6)* 市场上现有的解决方案未经调整无法满足 *contracting authority* 的需求；

该订单包括创建一种前所未有、极其复杂且全面的信息技术解决方案，涉及集成以下区域的组件：

> *1.* 对海量信息资源的访问安全；*2.* 为来自波兰各地科研机构的不同大学的大量用户实现访问联邦化；*3.* 将高性能计算解决方案与通过云计算模式同时提供这些资源相结合；*4.* 将云存储和对象存储解决方案与云计算相结合；*5.* 将批处理排队系统（用于高性能计算）与交互式访问、容器化和虚拟化相集成。

尽管上述每个组件和技术都分别存在于市场上，但据我们所知，在波兰或欧洲市场上，还没有任何一个安装部署案例能够同时满足所有这些要求。

整个解决方案属于所谓的“定制化”方案，必须与 *ICM* 多样化的计算机和超级计算硬件紧密匹配。这些硬件包括约 *600* 台服务器、采用各种技术的磁盘阵列、不同品牌且总容量约为 *25PB* 的存储设备，以及各种文件系统——从面向对象的文件系统到高性能 *Lustre* 系统。网络解决方案也需要升级，包括服务器内部以及局域网/广域网网络，同时还要提供网络设备，并使所有设施适应用户截然不同的需求。还应指出，该解决方案应尽可能基于现有硬件，仅最低限度地使用新的硬件组件。我们预计，两个集群中的互连网络需要升级。市场上没有能够提供端到端解决方案来应对像 *ICM* 这样高性能计算（*HPC*）中心的特殊需求，并达到这种复杂程度的产品。

该项目不仅包括购买组件，更重要的是进行集成和重新编程，并创建专业化服务。

例如，存储供应商（*NetApp*、*Panasas*、*IBM*、*HPE*、*DDN*、*Western Digital* 等）都有各自针对特定硬件的解决方案，而 *ICM* 则使用所有这些供应商提供的存储硬件——但这些硬件被存储孤岛之间厚重的“墙”隔开。全球只有少数公司能够提供所谓“硬件无关”的数据管理软件，也就是说，这类软件独立于硬件供应商——而这正是我们在存储领域所需要的集成解决方案。

所需解决方案的创新示例包括：
- 波兰目前没有能够以交互方式提供超级计算资源的解决方案。唯一可行的访问方式是通过 *SLURM*、*LSF Platform*、*PBS Pro* 等队列系统。- 将高性能文件系统（例如 *Lustre*、*Spectrum*）与基于对象的文件系统相结合的云存储，以及单一云计算- 将高性能计算资源与使用 *Docker/Singularity/* 其他容器化系统创建和运行作业的能力相结合

波兰不存在此类解决方案。

该项目不仅包括重建，还包括增加许多不同的元素、软件、进行整合和所谓的升级，即现代化改造。例如，*Topola* 和 *Enigma* 集群必须升级其互连网络（内部网络）。这两个集群均由华为制造，但互连网络产品由 *Mellanox*（现为英伟达）提供，因此解决方案必须由 *Mellanox* 和华为共同提供。

该解决方案的创新不会局限于各个要素的集成。

我们将要求该解决方案能够提供量子计算机模拟器的使用权限。也就是说，解决方案提供方必须拥有此类模拟器，或与能够提供此类模拟器的公司合作。需要强调的是，波兰没有此类供应商，而欧洲也只有少数几家。

在合同履行过程中，承包商还必须提供额外的硬件组件。例如，*Topola*和*Enigma*集群肯定需要互连组件（光纤、交换机、网卡、网络软件）。

根据我们对高性能计算、存储、网络等高端市场的综合了解（高性能计算、网络、存储、云解决方案、系统管理、文件系统、存储系统、容器化、排队系统等），没有任何供应商能够交付我们订单中规定的全部组件。我们预计，即使是该市场中实力最强的企业，也将不得不组建由数家分包供应商和分包承包商构成的联合体。

*7)* 工程、供应或服务包含设计或创新解决方案；

服务包含设计和创新解决方案，将目前尚无集成解决方案的各项功能结合起来。由于该解决方案具有创新性，因此有必要在设计前进行周密规划，并与承包商密切合作，以其提供的组件为基础，同时考虑 *ICM* 的硬件。我们强调，在当前的 *ICM* 项目中，整个解决方案具有很大的复杂性，实施难度极高，因为该方案必须包含许多新元素，同时还必须适配现有基础设施。

*8.* 因所涉事项的性质、复杂性或法律或财务条件特殊，或因工程、供应或服务所附带的风险，未经过事先谈判不得授予合同；

由于技术原因，该订单要求与承包商进行事先谈判：必须确保交付的模块彼此之间，以及与 *ICM* 硬件在硬件、软件和功能层面完全兼容，从而确保与外部系统充分集成。*ICM* 还保留根据自身要求和需求制定完整集成解决方案的权利；这些要求和需求将在对话过程中确定，且 *ICM* 有权不接受供应商通常提供的现成、开箱即用解决方案。

*9)* 如果 *contracting authority* 无法通过参照特定标准、欧洲技术评定（第*30*条第*1*款第*2(c)*点所述的合同标的说明）、通用技术规范（第*30*条第*1*款第*2(d)*点所述的合同标的说明）或技术参考文件，足够准确地描述合同标的。

显然，合同的组成部分不会是现成的解决方案。阅读资格标准及其他文件的任何人，都应当是被定义为“本领域技术人员”的人。对本项目所涉主题具有广泛知识的专家，完全有能力评估该项目，并能确保供应商有能力承担此类项目。

无法根据第*30*条第*1*款第（*d*）项第*2*点（*d*）描述合同标的，因为该订单涉及数十个要素的集成，而每个要素均为非标准要素。

鉴于合同在技术和概念方面的复杂性，以及许多区域缺乏适当标准，目前无法根据《公共采购法》第*30*条第*1*款第（*2*）（*d*）项所述技术和法律规定描述合同标的，即通用技术规范；通用技术规范是指根据《欧盟第*1025/2012*号条例》第*13*条和第*14*条（涉及《民法典》相关规定）所定义的信息通信技术产品领域的技术规范。该条例由欧洲议会和理事会于*2012*年*10*月*25*日通过，内容涉及欧洲标准化，并修订理事会指令*89/686/EEC*和*93/15/EEC*，以及欧洲议会和理事会指令*94/9/EC*、*94/25/EC*、*95/16/EC*、*97/23/EC*、*98/34/EC*、*2004/22/EC*、*2007/23/EC*、*2009/23/EC*和*2009/105/EC*；同时废止理事会决定*87/95/EEC*和欧洲议会及理事会第*1673/2006/EC*号决定（《欧盟官方公报》*L 316*，*2012*年*11*月*14*日，第*12*页）。

竞争性对话参与方的遴选标准已制定，并载于附录1。

竞争性对话公告于2021年1月发布。三家候选方于1月底选定；从2月初至7月，我们每周与三家潜在供应商分别安排一至两小时的会话。会话结束后，我们会进行内部讨论，并在当周稍晚时候形成新一轮问题和事项，传达给对话参与方。这一过程持续至2021年7月初。

VII. 教育社区、学者、学生和大学官员

ICM于

2018 2021 -期

[30] 2021

间主办了四届“Supercomputing Frontiers Europe”系列会议。该届会议设有云计算和高性能计算特别会场，并邀请了来自UberCloud、Rescale、Ronin和阿贡国家实验室的演讲嘉宾 [31-35]。ICM还组织了“ICM计算机与计算科学虚拟研讨会”系列，共二十一场。该系列的一场特别讲座由加州大学圣迭戈分校学术技术服务信息技术服务部高级主任兼SATO Valerie E. Polichar主讲，题为“构想技术愿景：规划新一代研究支持” [36] ，总结了在制定和实施“UCSD数字大学蓝图”之后，UCSD十多年来的经验 [37,38]。ICM试图效仿这一过程，在UCSD实现该经验中所确立的目标，并将计算资源拓展到远超传统高性能计算用户的范围。

## VIII. Final Specification 的 Solution

文件“合同标的说明”是本次函件的核心内容，全文收录于附录 2。

A-CHOICeM 项目遗憾的是始终未能实现。我认为，波兰及其他地区的学术中心有权了解这一耗时且严谨的过程及其成果，这符合公众利益；即使这些成果仅仅是积累的智慧与经验，以及对高性能计算云架构当前技术水平的了解。

## IX. 结onclusion 与 课essons 课earned

A-CHOICeM 项目始终未能完成。我们的 ICM 团队完成了技术规范文档，这是本报告的重要组成部分；但作者于 2021年 年底辞去华沙大学的职务后，该项目便被中止。

在两年的深入参与这一项目及其他工作的过程中，我们看到许多想法在各种项目和解决方案中得以实现——我们走在了正确的道路上。希望本文介绍的经验、规范和文档能够为其他学术机构和高性能计算中心提供一些帮助，并且即使对于学术云服务的公有云提供商和云技术开发者而言，也能证明具有实用价值。

APPENDIX 1

表格 I. 竞争性对话的遴选标准。

| # | 资格标准 | 最低条件 | 标准权重 |
|---|---|---|---|
| 1 | 在高性能计算环境中的安装经验<br>面向学术界的环境，科学或<br>研究客户，国家实验室及其<br>自有高性能计算中心（0-10）分 | 在高性能计算中心内至少实施两项安装<br>面向学术界的环境，科学或研究<br>客户，例如大学，学术高性能计算中心，<br>研究机构、拥有自身<br>高性能计算中心 | 10 |
| 2 | 云计算安装经验（0-<br>10分） | 至少实施过一次云计算<br>规模超过600台服务器的安装 | 10 |
| 3 | 使用对象存储安装云存储的经验，<br>总容量至少为10 PB（0-10分） | 至少实施一套用于数据<br>存储的云部署，使用对象存储，原始容量至少为 10 PB | 10 |
| 4 | 至少实施一套多云<br>部署（包括至少两个公有<br>云） | 至少实施一套多云部署<br>（包括至少两个公有云） | 10 |
| 5 | 将云访问与<br>联合身份验证集成的经验（0–10 分） | 至少实施过一次可访问<br>采用联合身份验证系统的云的部署 | 10 |
| 6 | 安装云计算的经验，包括<br>与 SLURM 排队系统集成（0-<br>10 分） | 至少实现一种云计算<br>安装，并与 SLURM 排队<br>系统 | 6 |
| 7 | 具有使用集成式安装的经验<br>访问门户，以便用户配置<br>虚拟机，容器，SaaS, 资源<br>（0-10 分） | 至少实现一种安装，并配备一个<br>允许用户配置的集成访问门户<br>虚拟机，容器，SaaS, 资源 | 10 |
| 8 | 存储数据安装体验<br>加密（0-10 分） | 至少实施过一次安装，具备<br>存储数据加密 | 5 |
| 9 | 具备安装经验，并与以下内容对接:<br>量子计算（0-10 分） | 至少实施过一次安装，并可访问<br>量子计算 | 5 |
| 10 | 具备与人工智能<br>工具的接口安装经验（0–10 分） | 至少实施一项安装，并配备<br>人工智能工具接口 | 8 |
| 11 | 拥有并运营一家研究实验室，专注于<br>云计算和高性能计算领域 | 雇用不少于50名从事研发工作的人员<br>岗位 | 6 |
| 12 | 专属团队（0-10分） | 为项目介绍至少4名关键人员，并附上<br>对其经验的描述 | 10 |
| | | | **100** |
| | **资格标准** | **准入条件** | |
| 13 | 具备所需的工具和组织及<br>技术手段，以确保妥善<br>履行合同 | 请说明拟采用的工具和方法论，以开展<br>项目管理 | – |
| 14 | 其公有<br>云服务尤其需要<br>这类组织必须具备的认证：SOC 1、SOC 2、SOC 3、CSA STAR、<br>ISO 9001，ISO 27001，ISO 27017，ISO 27018 | 请列出承包商持有的所有认证<br>从此列表中 | – |
| 15 | 管理委员会成员的无犯罪记录证明<br>（现任） | | – |
| 16 | 国家法院登记册副本（标准版） | | – |
| 17 | 财务报告/损益表<br>最近 3 年（0-10 分） | 最低收入达到 8000 万波兰兹罗提 | – |

APPENDIX 2

DES合同标UBJECT 的ATTER 的合ON 同

本订单涉及利用现有的 ICM 基础设施实施云计算 [ （以下简称“项目”）]，并包括：

a. 实施必要的软件解决方案 b. 补充并升级现有 ICM 设备 c. 编制文档 d. 提供培训 e. 全生命周期支持

合同标的说明

第*1*部分：项目目的

第*2*部分：要求

1. 供系统管理员使用的用户访问与管理门户。 2. 安全 3. 虚拟化、容器化、多云和混合云功能 4. 应用子门户（区域） 5. 架构i 云存储 ii 网络架构iii 云架构与服务iv 集成 6. 文档7. 培训8. 支持与保修 9. 实施与付款计划 10. 订单收集方式

（附件*1*——采购方可用基础设施）

**第ART 1： 第URPOSE 项目**

本项目旨在为华沙和波兰的学术及科研界创建一个易于访问且易于使用的学术云计算系统 ICM（ A-CHOICeM）。本订单包括创建一个前所未有、极其复杂且全面的信息技术解决方案，涉及集成以下区域的组件：

1. 便捷、易用且安全地访问海量 IT 资源，

2. 为来自波兰各地科研机构和大学的大量用户实现访问联邦化，3. 提供高性能计算解决方案，同时以云计算模式提供这些资源（包括虚拟机和容器化），4. 将云存储和对象存储解决方案与云计算相结合，并集成批处理排队系统（用于高性能计算），支持交互式访问、容器化和虚拟化。5. 存储——集成各种类型的数据存储资源，同时向用户开放 A-CHOICeM 云。

A-CHOICeM 云的预期功能：

1. 混合性 - 通过将该解决方案与 SLURM 排队系统集成，该云能够同样便捷地创建虚拟计算资源，以及访问大规模高性能计算资源。 2. 多云 - A-CHOICeM 用户门户可访问 Azure、AWS、Google Cloud、IBM、GreenLake 等公有云。 3. 云存储 - 可通过门户使用，支持轻松创建存储资源，并根据特定使用策略自行管理这些数据。 4. 容器化 - 能够使用和创建 Docker、Singularity 等常用容器。 5. 为波兰大学和研究机构的科研人员、教师及学生提供通用且便捷的访问。 6. 访问安全以及存储数据和用户账户的安全。 7. 针对常见研究领域精心设计且可直接使用的子门户，例如：8. 量子化学、材料科学、生物信息学、软件开发和数学建模（例如 R、Python、Julia、Mathematica）、结构工程、计算流体力学。子门户必须配备这些领域中最常用的代码、工具和框架，或已准备好与其集成。 9. A-CHOICeM 云可访问人工智能（AI）和量子计算领域的先进门户/服务及工具（包括公开开放的，以及在可能情况下的专有门户/服务及工具）。10. A-C HOICeM 云采用基于开放许可证的组件构建，并支持进一步扩展和扩容，以连接更多本地和远程硬件资源及其他云。

A-CHOICeM 旨在提供一种解决方案，使用户能够访问种类丰富得多的 ICM 计算资源，并帮助 ICM 满足科学研究和教育计算以及数据存储方面的全部需求

，并将学术界和科学界的资源开放给数量显著更多的用户。

本项目旨在创建一种解决方案，实现对多样化基础设施的便捷访问和自主管理，涵盖基于单个计算核心的单台虚拟机、ICM 提供的超级计算机计算能力，以及多个公有云中提供的各种资源。

**PART 2: REQUIREMENTS**

Academic Computing Cloud ICM——“ACHOICeM”的设计与实施订单包括交付一种技术解决方案，该方案由满足以下要求的组件构成：

**2.1 系统管理员用户访问与管理门户。**

A-CHOICeM 访问门户的首要要求是为用户和系统管理员提供易用性与便利性。

必须为学术和科研界人员提供联合访问，既包括来自华沙的人员（即 UW、PW、WUM、WAT、SGH、SGGW 等），也包括来自波兰各地的人员。

用户与使用情况

云用户包括：

1. 经授权的研究人员或学生，他们使用云提供的计算资源和软件 [**最终用户**]（具有不同级别的授权）；以及 2. 具有不同权限级别、负责管理这些资源的 ICM 管理员 [**管理员**]。

此外，管理员可根据权限级别，为特定最终用户或用户组定义可用资源和服务目录。权限可以组合成权限组，并分配给用户或用户组。

根据管理员分配的角色，最终用户可能拥有不同级别的权限，可共享资源或为其他用户分配角色。例如，代表某个单位的大学、院系或研究所用户组负责人，有权共享分配给该单位的计算资源和数据存储资源。

此外，系统中可以为项目、资源和用户组设定特定的有效期限。例如，由管理员指定的主管可以根据适当制定的访问策略，向参加特定大学课程的一组学生授予访问虚拟机和容器化计算环境的权限，其中包括特定软件。

**门户功能及其访问方式**

最终用户将通过以下方式访问云资源：

- **网页界面**（管理用户的云资源、访问权限并监控资源使用情况），
- **API**
- **命令行界面**（SSH）——直接访问用户的虚拟机，使授权用户能够在管理员分配的资源范围内自主配置虚拟机，并监控这些资源的使用情况。
- **S3 协议** ——访问磁盘资源

用户界面的最低必备功能包括：

1. 用户身份验证；2. 用户档案管理；3. 已分配资源及其使用程度的信息；4. 与工作人员沟通（报告问题或提出疑问）；5. 使用用户自定义参数创建/激活/删除虚拟机；6. 创建/删除虚拟网络；7. 分配/释放虚拟机的 IP 地址；8. 为虚拟机分配/释放卷；9. 创建 S3 资源；10. 创建/激活/删除容器；11. 通过图形界面访问各种文件系统中的磁盘资源（例如 S3、Lustre）；12. 通过图形界面访问高性能计算系统；13. 通过单点登录访问集成的公有云资源；14. 访问代码仓库（用于构建容器的配方）；15. 容器镜像仓库（通用仓库和用户组专用私有仓库）；16. 访问一项服务，该服务可根据代码仓库中的配方构建并存储容器；17. 用户界面符合 WCAG 2.0 标准。

网页界面软件的源代码及相应许可证必须提供给 contracting authority，并且必须妥善编制文档，以便 contracting authority 日后能够自行修改。

网页界面必须适用于移动设备和台式计算机。

管理员可用的功能：

1. 所有终端用户界面功能；2. 用户、角色、机构、项目、映射，以及用户组/类别/层级的管理；3. 软件管理；4. 策略和云资源的管理（计算核心、存储、网络、每位用户的限制、每个用户组的限制以及配置周期）；5. 计量——统计用户对资源的消耗；6. 自动生成报告，汇总用户、组和机构层面的资源消耗；7. 监控云资源；8. 管理终端用户类别、账户到期时间和资源限制；9. 许可证管理。

## 2.2. 安全

ICM 云基础设施必须满足以下云系统和存储数据的安全要求：

1. 只有授权用户才能访问云资源。2. 用户身份验证基于单点登录机制，并采用双因素认证。所实施的解决方案必须支持与各种外部用户数据源集成（PIONIER.id、CAS、授权机构的电子邮件地址）。所采用的双因素认证机制应尽可能通用，让尽可能多的用户都能使用，且无需在用户端安装其他应用程序或使用专用硬件设备（令牌）（优先考虑例如通过短信发送一次性验证码，和/或发送至用户账户中已定义并验证的电子邮件地址的一次性验证码）。同时，双因素认证机制必须能够与所有已实施的用户授权机制和来源协同工作。

3. 资源访问的分层隔离 此处所说的分层访问隔离，是指能够配置系统资源和数据的权限，使得：

- 确保各租户之间的资源访问完全隔离，
- 在分配给租户的资源范围内，可以配置特定用户组对资源的访问隔离，
- 在分配给用户组的资源范围内，可以配置特定用户对资源的访问隔离，
- 在这两种情况下（租户内和组内），都必须能够共享某些资源。

4. 云计算基础设施在互联网上提供的服务必须经过加密。5. 基础设施应至少具备以下事件监控功能：

1. 可用性，2. 资源使用情况，3. 有助于预测故障可能性的参数（例如 *S.M.A.R.T*），4. 由软件错误导致的活动（例如应用程序意外中断）。6. 集中式日志记录的用户活动至少应包括以下内容：

- 启用/禁用/创建基础设施组件（虚拟机、容器、虚拟网络、分配的磁盘资源等）
- 分配的 *IP* 地址，
- 已分配许可证。

7. 环境必须支持更新所用的软件（OpenStack、操作系统、虚拟机的系统映像、应用软件等）（参见第 11 点）。

8. 删除虚拟机时，基础设施必须自动清零空闲存储空间和空闲内存。此过程应完全自动完成，无需用户或管理员进行任何干预。9. 基础设施必须提供不同级别的备份功能，至少包括以下内容：

- 整个环境配置的副本（支持回滚），
- 门户及其配置的副本，
- 用户数据的副本，
- 在用户/租户侧启用已分配存储空间的备份。

10. 所有直接影响业务连续性和可用性的基础设施组件都必须实现冗余。

## 2.3 虚拟化、容器化以及多云和混合云功能

1. 门户需要通过访问虚拟机、容器化环境和由 SLURM 管理的 HPC 集群，充分利用可用资源。

对于虚拟机和容器，该解决方案还将支持定义可供使用的资源总预算。具体而言，该解决方案必须允许定义：

- 用户/组可用的资源小时数总量（尤其是 *CPU* 小时/*GPU*小时）• 已分配预算的有效期• 指定资源池的负责人（项目经理）

该解决方案将通过有文档说明且安全的 REST API，提供对完整计费数据（包括预算）的访问。既可以访问用户自己的数据（用户本人可以查看其预算以及

在指定时期内使用了多少资源），也可以访问所有用户的完整计费数据（需具备相应权限）。外部系统也必须能够通过同一 API 层级全面控制可用预算（需具备相应权限）。该解决方案还必须能够配置所有用户均可使用的基本预算（“免费层”）。该解决方案必须收集并提供预算使用情况数据，数据更新频率至少为每小时一次。预算即将用尽时，必须通知用户。

该解决方案还将提供限额支持（按用户、按用户组），并对所有（虚拟）机器进行汇总：

• 同时运行的机器数量• 已占用核心总数• 单台机器使用的 *CPU/GPU*/核心数量• 已使用的内存容量• 公有 *IP* 地址• 存储资源（*GB*、*IOPS*）

计费系统将收集在中央容器化集群上运行的虚拟机和容器的数据。

用户提供的所有配置项（包括 Dockerfile、Ansible 脚本，或在容器中运行或由领域应用运行的脚本等）都应可从集成的私有 Git 仓库下载或存储；该仓库与单点登录集成。

凡是需要存储虚拟机镜像或容器的地方，该解决方案都将提供适当的机制（例如镜像仓库），并使用客户指定的 S3 资源作为存储位置

2. 门户应支持通过三种不同方式访问资源和应用程序：

A. 虚拟机访问

作为面向用户的门户所提供服务的一部分，最终用户必须能够使用目录中提供的镜像、从公共存储库导入的镜像或自行提供的镜像来运行虚拟机。图书馆中的预定义镜像必须包括当前稳定版的 Ubuntu、CentOS、Rocky Linux、Debian、Alpine，以及 CentOS 7。此外，门户还应允许创建虚拟机的快照和副本。用户提供或导入的镜像将计入该用户占用的总存储空间

门户必须向用户提供：

• 通过 *SSH* 访问您自己的虚拟机所需的参数

• 通过 *Web* 浏览器使用 *VNC* 访问您自己的虚拟机（虚拟桌面）• 允许为预定义的虚拟机指定参数• 允许控制虚拟机：启动、关机、重置

• 通知（通过电子邮件）用户即将到来的时间限制或资源强制释放。

B. 容器集群访问

该解决方案将创建一个集中管理的容器化集群，允许用户运行来自本地存储库、从外部镜像仓库导入的镜像，并通过指定公共 GIT 存储库或用户资源来构建镜像。本地镜像仓库（例如 Harbor）应能区分镜像

由 ICM 提供并支持。来自外部代码仓库的镜像必须通过本地存储库下载，并在可供运行前进行自动化安全检查。在中央 Kubernetes 集群中运行的容器不得来自中央镜像仓库以外的来源，或通过中央镜像仓库代理访问的外部镜像仓库。用户应能够选择使用自己的 Kubernetes 集群（运行在自有预算购置的虚拟机上），或使用共享资源来运行容器。

C. 访问 HPC 集群

访问 Okeanos 集群和 Poplar 集群中由 SLURM 工作负载管理器管理的资源。在此解决方案中，所选配置使用用户提供的集群访问参数。访问范围、限制及计费应由一个独立系统处理。交付的解决方案应支持将访问 HPC 集群作为流水线中的一个环节：在集群上运行的作业将结果传递到用户的虚拟机，并在该虚拟机上对结果进行受监管的后处理。

3. 应用使用模型

无论以何种方式访问资源，对于目录中定义的每个应用程序，我们都区分两种基本使用模式之一：任务（批处理模式）和服务（交互式应用）。同一软件可能以不同变体提供，但会将其视为两个应用程序（例如，Julia 脚本（任务）和 Jupyter Notebook）。

任务应用/批处理模式

任务型应用程序用于处理某些预定义任务，并在任务执行完毕或失败前持续消耗资源。预定义任务应用程序由

用户从目录中选择，或基于通用模板（例如 Python 脚本应用程序）自行创建。用户必须能够与其他用户（群组成员、指定用户或整个中心）共享自己的应用程序。用户需要配置运行任务所需的一组参数，包括指明要传输到任务中的磁盘资源（文件、目录）。随后，任务将使用应用程序定义中指定的资源运行（虚拟机、中心环境或自有环境中的容器、高性能计算）。作为用户自有项目示例和模板的通用应用程序，必须包含任务型应用程序在所有可用资源类型上运行的示例。用户可以将任务型应用程序组合成流水线，其中一个任务的输出（例如目录、tar 归档文件）会作为输入传递给流水线中的下一个任务，并定义依赖关系和重启机制，类似于 GitLab CI/CD 流水线、Argo Workflows。作为此类工作流的一种极端形式，应当能够附加服务型应用程序（例如用于数据预处理和后处理的虚拟桌面）。在初始安装时，必须提供应用程序，以便使用本文档后续介绍的软件包运行任务和分析。

在任务（批处理）模式下，用户应能够从感兴趣的类别中选择所需的应用程序，然后将一组批处理数据上传到正在运行的程序。任务执行期间，用户应能够控制正在运行的应用程序的状态，对其执行基本操作（包括取消任务），查看程序的标准输出，并查看生成的输出文件内容。整个过程应尽可能减少对用户的干扰；不过，高级用户应能轻松使用更广泛的干预功能，尤其是能够：

• 自行定义目标节点上运行的队列系统如何调用应用程序，即直接在命令行中提供程序名称、选项和调用参数；• 控制标准输出，即通过流水线机制处理生成的数据，尤其是使用 *grep* 工具、*sed* 流编辑器以及符合 *POSIX* 标准的系统标准环境中提供的其他实用程序筛选所需模式；

• 请求在应用程序正常退出或未能正常退出后执行某项操作，尤其是向指定地址发送电子邮件，或执行以 *shell* 语言（*Bash*）编写的自定义脚本。

任务/任务模式还必须能够轻松地通过拖放方式定义高级工作流，包括多个应用程序的执行顺序，以及它们对先前任务生成的数据的使用方式、基于应用程序运行结果定义条件指令等。

使用 Quantum Espresso 应用程序进行密度泛函理论（DFT）电子结构计算的工作流示例。以下每个步骤都需要使用上一步生成的数据：

• 使用*“pw.x”*程序，通过收敛结束的自洽循环（*SCF*）——其数据和参数由用户在批处理文件中设定。系统应能够验证是否已达到收敛，从而判断是否可以进入下一阶段（由用户编写的 *shell* 语言条件指令）。

• 使用自洽场（*SCF*）循环生成的数据进行非自洽循环。计算倒易网络中给定一组*“k”*点对应的 *Kohn-Sham* 态。

• 结果后处理。创建工作流时，用户需定义要对哪些文件进行后处理，以及如何处理：使用 *shell*（*Bash*）指令或脚本，其中包含程序调用、选项、参数、批处理数据和结果数据，以及 *I/O* 重定向和流水线。

另一种典型的使用场景是：对批量数据（由用户上传或从指定的磁盘资源下载）运行应用程序，完成后再以图形模式运行交互式应用，进一步处理生成的数据。

创建并连接工作流任务的机制应支持基于 Bash shell 语言的多种配置选项。在每个步骤之间，应能够定义条件指令（以及较大的脚本），并据此决定是否继续执行。其目标是尽可能实现工作自动化，并在一次性定义任务后执行多个相互依赖的任务。系统应允许创建此类任务流水线的模板，以便日后使用和修改；也应支持使用其他常见的工作流系统，例如 Kepler（另见第 5.4 节：集成）。

4. 服务 / 交互式应用程序

此类应用程序从启动到用户终止期间都会消耗资源。此类应用程序必须能够通过互联网共享自身资源，或限制在内部网络中（为其用户提供连接到云的共享 VPN）。提供应用程序访问权限的基本方式包括远程桌面访问（VNC、WebVNC）和端口访问，例如 HTTP/S。交互式应用程序以虚拟机或容器的形式运行。一个应用程序可以运行多个虚拟机，例如，预先配置的 Kubernetes 集群，其中包含内部预先准备好的 JupyterHub 环境。根据预先准备的配置，应用程序应能够使用用户的磁盘资源，或与这些资源安全隔离。由于此应用程序的终止

与用户操作相关，因此有必要为每个应用程序定义其自动终止时间，并了解正在运行且消耗其资源的应用程序。拟议的解决方案必须包括以下服务/交互式应用程序，这些应用程序专为在私有微型集群上运行而设计，并会自动为用户配置：

- *Jupyter/Jupyter Lab/Jupyter Hub*，可选择支持用户文件操作。预定义解决方案应包含一个专用于学生/受训者工作的版本（通过带有公共 *IP* 的哈希链接进行分享）。应用程序应安装支持以下语言的 *Jupyter* 内核：*Python*、*Julia* 和 *R*。还必须允许选择容器工作节点（例如，带有独立 *Java* 环境的节点），以便在其中运行笔记本（*Jupyter Hub* 的标准选项），
- *RStudio*，支持用户文件操作
- *BinderHub*，支持使用与其兼容的公共代码仓库。

此外，该解决方案必须包含适用于以完全图形模式运行预处理/后处理任务的虚拟机镜像，并基于当前稳定版 Ubuntu 和 CentOS 操作系统。用户应能够通过 VNC 访问基于该镜像运行的应用程序：

- *ParaView*
- *ANSYS*、*Matlab* 和 *Mathematica* 软件包，使用 *ICM* 提供的许可证或用户指定的许可证
- 一个简洁的 *GNOME* 桌面，可访问用户文件并能够安装软件

上述应用程序可以组合在一起，作为一个“虚拟桌面”提供。

用户还应能够直接从门户运行在模拟文本终端中运行的应用程序（无需另行建立独立的 SSH 会话）。

5. 应用程序使用场景

为更好地从最终用户的角度说明这些要求，A-CHOICeM 必须满足以下使用场景：

A. 使用 JupyterHub 进行培训

讲师（注册用户）希望通过将链接指向以下位置，分享包含为培训准备的材料的交互式 Jupyter 笔记本：

- 私有/公共代码仓库。
- 私有/公共 *Docker/Singularity* 镜像。

讲师可以控制每位用户可共享的计算资源数量（核心数、内存、工作区容量、任务/会话的最长时长）以及用户的最大数量。受训者消耗的资源计入讲师/受训者项目（取决于所选选项）。受训者不必是 A-CHOICeM 用户，讲师可以选择“通过已知链接授权”

B. 使用虚拟机（VM）镜像进行培训

讲师准备了安装有所需软件和批处理文件的虚拟机镜像，用于开展课程。受训者可以克隆讲师准备的虚拟机镜像，并在自己的账户中运行该镜像。受训者消耗的资源将计入讲师/受训者项目（取决于所选选项）。

C. 开放培训

开放培训是指无需用户注册。讲师提供魔法链接（每位参与者一个或所有参与者共用一个）后，参与者即可启动 JupyterHub 或 BinderHub 会话。创建活动（培训）时，讲师会设置“魔法链接”的到期日期。

D. 文件共享

属于同一项目的用户可以通过预定义目录共享文件。

E. 现状

用户仍可像以前一样通过 SSH 登录 HPC 集群，并向 SLURM 队列系统提交作业。

**2.4. 应用子门户（区域）**

2.4.1 门户面向最终用户的功能将包括在预定义的一组应用程序中运行计算任务。可用的应用程序应按代表不同研究领域的组进行分类。类别和应用程序的便捷管理（添加、删除、更新）以及运行方式的管理，应由门户管理员负责。以下列出了所提供服务中应包含的主要类别和应用程序（某些应用程序可能属于多个类别）。2.4.2 应用程序的可用性是指能够通过上述至少一种访问方式运行该应用程序。

2.4.3 研究区域和应用程序清单应被视为示例性目标产品。门户应实现若干典型且文档完善的服务交付模式，其中包括

未来复制和开发这些应用程序的可能性。具体而言，该解决方案必须提供现成的服务，以运行以下应用程序（另请参阅应用使用模型）：

- Quantum Espresso
- Gromacs
- Mathematica
- Python / TensorFlow / PyTorch
- ANSYS Fluent
- C/C++/Fortran 编程环境（GNU、Intel、数学库）

2.4.4 目标研究领域和应用程序：

材料计算 /物理学 / 量子化学
- *VASP*
- *Wannier*
- *SIESTA*
- *ORCA*
- *NWChem*
- *NAMD*
- *LAMMPS*
- *Gromacs*
- *Gaussian*
- *Quantum Espresso* •

*Elk*
- *Dalton*
- *CP2K*
- *Abinit*
- *CASTEP*

分子动力学
- *Gromacs*
- *NAMD*
- *LAMMPS*
- *Charmm*
- *Desmond*

计算机代数 / 其他
- *Mathematica*
- *Julia*
- *GNU Octave*
- *Matlab*
- *R*
- *Python* •

*Anaconda*

计算流体力学：计算流体力学
- *ANSYS Fluent*
- *OpenFOAM* 有两个版本，均拥有众多用户：- *OpenFOAM* – [39] *(ESI-OpenCFD)* 的最新版本为 *v2106*。- *OpenFOAM* – [39] *(Foundation)*——最新版本为 *v9*，以及第三方软件包。

机器学习
- *TensorFlow*
- *PyTorch*

编程语言
- *Julia*
- *Python* •

*Go*
- *R*
- *Java*
- *GNU*：*C/C++/Fortran*
- *Intel*：*C/C++/Fortran*

软件开发、编译器和库
- *GCC*：*GNU* 编译器套件 • *Intel*（*C/C++/Fortran*）开发工具链（*MKL* 等）• *BLAS*、*LAPACK*、*ScaLAPACK*、*GSL*
- *OpenMPI*、*MPICH*、*MVAPICH* •

英伟达的奇迹
- *LLVM*（*Clang*、*Flang*）
- *CMake*
- *Spack*

工程
- *LS-Dyna*

其他
- *Garuda*
- *Taxila*

除了现有基础设施中可用的应用程序外，门户还应支持未来集成外部服务提供商的可能性，例如

- Wolfram 企业私有云 - 在本地运行，作为私有云的一部分，或接入由外部提供商支持的服务。

- 量子计算 - 量子计算机和模拟器，例如：- IBM Q（可访问 IBM 量子网络）- Quantum Inspire [40] - Amazon Braket - Forge（D-Wave、Google、IBM）

- 人工智能 - 深度学习（例如 AWS）

### 2.5 架构

整个解决方案属于所谓的“定制”（定制化）方案，必须严格适配 ICM 拥有的各类计算机和超级计算设备，包括约 600 台服务器、采用不同技术和品牌、容量约为 25 PB 的磁盘阵列，以及各种文件系统 -

从面向对象转变为高性能 Lustre 系统。

拟议的解决方案应考虑最大限度地利用附件1中列出的资源，如OPZ所述。未来，这可能需要~~制定分~~阶段计划：第一阶段在本文档中说明，之后可在后续阶段将其他资源转化为云服务，继续推动现有资源向云服务转型。

2.5.1 ICM现有基础设施的规模与使用

ICM云计算将基于客户现有的基础设施（计算节点和存储）向用户提供虚拟计算节点。云应能优化利用订购方在附件1和OPZ中所述的所有现有设备。

系统应支持未来通过增加设备（服务器、交换机、存储资源）来扩展资源。

实施文档应包含与交付解决方案兼容且支持无缝扩展的硬件要求。

应预先考虑并记录未来增加资源（节点的数量和类型、磁盘空间）以及扩大使用规模（用户数量）的可能性，至少支持2000个节点。在当前阶段，解决方案应使用现有的Enigma硬件（用于云计算和存储），并支持连接现有的Okeanos系统资源（通过SLURM）。

2.5.2 云存储

承包商应基于Enigma集群的单个模块实施数据存储资源共享系统（“云存储”）。云存储系统应采用分层存储结构进行组织，并能够根据管理员定义的策略自动进行数据迁移。文件系统必须根据各云节点的权限进行集成并可供其使用。该系统必须与雇主的其他磁盘资源（第5.4点）完全集成。

1. 解决方案的技术条件 1. 根据第1节所述要求访问门户。

2. 基于面向对象的数据存储技术，兼容Amazon S3 API 3. 数据冗余和安全：- 可根据对象大小选择复制或纠删码 - 支持为不同对象采用不同的纠删码配置文件 - 自动或按需进行集群再平衡，例如添加新磁盘或移除磁盘后 - 在存储集群中添加和移除磁盘及服务器时，不影响数据共享 - 集群运行期间可进行不间断升级（滚动升级） - 能够为存储数据创建策略（生命周期策略、一次写入多次读取等） 4. 数据应可通过以下方式访问：- S3 - 网络文件系统 - Web门户 - Share（相当于Amazon Simple Cloud Storage）- Samba 5. 可选择基本存储块大小（数据块） 6. 可扩展性：- 集群中的所有节点都参与数据处理，确保性能与参与处理的节点数量成正比 - 硬件资源无需同质化 7. 授权和身份验证 - 能够与现有身份管理系统集成，例如 LDAP、AD、Linux PAM、基于令牌的 S3 和单点登录（SAML 2.0、OpenID连接器）- 该解决方案还应支持使用 REST API 管理资源。- 能够委派权限管理。8. 资源消耗的计量和控制 - 创建用户资源使用情况的详细报告 - 设置用户最大磁盘空间使用量（配额）上限，以及数据传输速率和存储期限限制。9. 该解决方案应能够使用 AES-256 加密数据（传输中和静态数据）。10. 该解决方案必须主动检查数据完整性，并修复检测到的任何问题。11. 通过 API、控制台（命令行界面）和 Web 进行管理。12. 永久许可证，可用于现有集群资源。

2.5.3 网络架构

云将部署在现有网络基础设施（1G 或 40G 连接）上，同时云软件必须支持与吞吐量更高（100G）的网络基础设施协同工作。

A-CHOICeM 云的网络基础设施必须以确保高效、可靠的连接为目标进行部署，为用户构建接入网络，并为云管理和存储资源访问构建完整的网络。

网络连接需求分为“数据网络”和“管理网络”。

数据网络旨在实现：

1. 云资源访问互联网；2. 云资源访问 UW ICM 网络；3. 访问存储资源。数据网络必须支持：

- 连接到单个云节点，吞吐量至少为 40G- 连接到 PIONIER 网络和互联网资源，容量至少为 100G- 创建虚拟网络，并从私有和公有地址池中动态或静态分配地址。- 通过分配的公有 IP 地址连接互联网。- 使用共享公有 IP 地址，通过 SNAT 连接访问互联网- 使用共享公有 IP 地址，通过 DNAT 连接从互联网访问- 对分配给用户的资源进行核算（尤其是共享 IP 地址的使用情况）- 扩展网络（通过增加地址池来扩充 IP 地址资源）。管理网络旨在实现：用于管理、监控和配置云资源的连接

管理网络必须支持云节点之间至少 1G 带宽的连接

此外，架构必须满足以下要求：

- DNAT 和 SNAT 功能必须在云节点的虚拟资源中实现。
- DNAT 功能必须由用户在已分配的限额内进行配置。
- 如果订购方在质保期内升级云网络基础设施，承包商有义务对云进行一次重新配置。• 该解决方案应能够定义软件定义网络（Neutron 软件定义网络）；计算节点应能够直接访问部分网络（二层网络），并通过专用网络节点间接访问外部网络。
- 云网络节点必须提供冗余。

2.5.4 云架构和服务

1. 云将基于 **OpenStack** 环境。整个系统应尽可能广泛地使用开源组件构建。如果使用的组件采用非开源许可证，则所提供的许可证应允许在交付之日起 3 年内免费升级（尤其是安全升级），并且即使缺乏支持，也应允许永久运行。2. OpenStack 控制器服务器（控制器节点）将部署在采用 HA 技术构建的虚拟化集群（底层云）上，该技术应支持服务器无缝迁移、从集群中添加和移除服务器，以及控制器的备份和恢复，同时不影响云解决方案的运行。3. 云部署机制应可复现（所交付的安装“场景”必须能够从头开始重新创建云的初始状态，即从裸机到可供使用的状态）。4. 控制器集群中应为支持相关服务的虚拟机预留空间（门户、模板存储库和配置、IdM、监控和日志记录等），并应据此规划该集群的规模。5. 正在构建的云环境应提供由以下内容组成的最低限度服务集：

- 在云基础设施中添加、移除和管理服务器（金属即服务）。• 用于在 IaaS 平台上管理实例、网络以及用户和项目访问权限的服务。
- 为虚拟机、物理机和容器提供存储组件（类型：映像、卷、S3 等）。支持多层级，例如：在不同类型、位置或可用性的磁盘上构建 S3。

- 符合 OpenStack 网络 API 的网络服务，可提供网络、子网、路由器、负载均衡器、VPN、防火墙和 IP 地址分配等管理功能。• 符合 OpenStack 身份 API 的集中式客户端身份验证、服务发现和分布式访问控制服务。

- 用于收集和监控系统使用信息的服务。该服务必须从 OpenStack 和其他云组件收集信息，并支持监控和计量资源使用情况：

- 用于计费用途• 用于管理用途

7. 中央系统日志收集服务。

- 用于管理配额（对分配的资源设定的限制）的服务，可在达到指定值时自动执行操作。

- 基于云的应用编排服务。

- 资源预留服务。

所有服务都应采用冗余架构，以确保高可用性。

云将使用 POZ 的 !"#$ &$ 中所述的现有存储基础设施进行部署。

此外，根据本采购订单中规定的功能需求，除上述组件外，可能还需要提供、安装和配置其他组件。承包商必须预见到并在本订单范围内完成所有必要附加组件的交付、安装和配置。须在报价中注明并说明这些活动。

2.5.5 集成

创建云解决方案的所有门户都应集成承包商实施的单点登录（SSO）系统，并同时支持多种身份验证机制，至少包括 PIONIER.Id、CAS UW、LDAP ICM（Keycloak），以及发送至电子邮箱的一次性密码。后一种机制必须在实施 SSO 系统期间交付，并允许编辑可接收密码的邮件域名列表。

2.5.5.1 与订购方的资源和 服务的集成：

门户必须允许访问由 SLURM 工作负载管理器（Poplar、Okeanos 系统）管理的资源。用户可通过门户提供的通用模板，连同批处理数据一起提交访问参数。门户应支持创建、修改和管理自定义模板，并提供便于根据正在运行的应用程序需求进行定制的功能，例如支持针对不同批处理数据运行多个程序实例。定义任务后，系统应自动将任务提交至队列系统，以便在其控制下执行。应用程序的使用机制（使用模型）详见第 3 节。门户的设计应支持以批处理模式（见第 3.1 节）和交互模式（见第 3.2 节）运行应用程序。不同工作流中运行的应用程序应能共享磁盘空间，以便进行后续处理（预处理/后处理）。解决方案应使用现有基础设施中的磁盘资源（S3 存储）。用户应能够跟踪正在运行的作业及其状态和参数，预览生成的输出数据，并在 SLURM 系统控制程序（如“scontrol”及其他程序）所提供的范围内，直接通过门户控制作业的执行。还应支持高级工具使用方式，允许用户输入自定义指令并将其转发至队列系统。

云存储集成要求

1. 该解决方案部署在现有硬件上，承包商将拥有一个完整的 Enigma模块可供使用。2. 存储云应与云解决方案的其他组成部分集成（使用相同的身份标识、身份验证机制和权限授予机制）。3. 希望将现有的 S3（Quantum）和 Lustre 解决方案集成起来，消除不同资源数据存储（Quantum/Enigma//Tetyda）之间的壁垒，例如使用 OneData/DataCore（前身为 Caringo）等数据提供商。4. 计算部分可以使用存储部分中的数据，反之亦然（例如，可以从存储部分提取数据，在生成结果后再将其放回存储部分）。5. 负责“计量/计费”的部分应实现集成，并涵盖计算和云部分。

5.4.2 与外部资源集成：

用户门户应允许访问量子计算、人工智能和机器学习门户、生物信息学，以及 Garuda、Taxila 和图计算（Urica XC、Trovares、Jena）等内容。

门户必须允许用户随意切换到热门公有云 Azure、AWS、IBM、Google a nd 其他内容。

**2.6 文档**

投标人将提供以下文档：1. 完整且可编辑的实施文档，用于说明服务器、网络、存储、服务和软件的配置。特别是，文档应包含网络拓扑、寻址、连接方案、已部署服务和已配置设备的配置文件、所用软件的版本以及密码等信息。2. 对解决方案中所用组件及其相互关系和关联的说明。3. 云的实施流程。4. 云的基本使用流程说明。至少应涵盖以下事项：

- 启用和停用云集群（包括底层云和上层云）的流程 • 备份与恢复流程 • 故障切换流程——在任一服务器或服务发生故障后恢复服务 • 回退流程——说明基础设施元素发生电源故障或硬件故障时应采取的措施

- 云软件更新流程 • 门户重新安装流程 • 更新 SSL 证书的流程 • 使用新服务器或交换机扩展集群的流程

- 从管理员的角度介绍云运维，例如添加新的身份提供商、更改服务策略、添加新网络、设置配额、建立用户层级并为本地管理员分配权限等。

- 介绍如何创建新的应用程序和服务模板并将其添加到目录。提供用于部署虚拟机、Kubernetes 集群和应用程序的现成且有文档说明的模板示例。 • 介绍监控系统和集中式运维日志服务，并说明与这些服务相关的基本操作（配置和使用，例如架构、备份、信息提取等）。
- 处理来自外部/内部 IP 地址的攻击的流程

- 账户遭到入侵时应遵循的程序（授权用户、管理员）

- 计费（REST API 描述、数据格式、导出、导入）。5. API 描述 6. 实用的云用户指南。介绍如何在云中执行基本操作，面向的对象不一定是技术人员（例如如何启动虚拟机或应用程序/服务）。7. 说明未来可能如何扩展的图表和描述。

**2.7 培训**

合同验收后的 3 个月内，承包商将为 contracting authority 的管理员提供培训：

- 云部署——讨论所使用的工具和组件、所做的假设、采用的方法、安装脚本的本地化，以及根据订购方的需求对其进行调整。 • 日常维护、基本操作、软件升级、紧急关机、提升、备份恢复、硬件更换、扩展、网络重新配置。 • 详细讨论系统和最终用户部分的网络架构（隧道、路由）。各组件之间的相互作用（网桥、协议）。 • 安装和更新用户应用程序、镜像、权限管理

投标文件将包含一份培训计划。

**2.8 保修与支持**

1. 根据合同提供的所有技术设备均应享有至少24个月的制造商保修，保修服务在安装地点提供，并确保设备发生故障时，自收到通知起48小时内更换为可正常运行的设备。2. 此外，承包商应自订单验收之日起，为所交付的软件向订购方提供为期24个月的保修。在保修期内，承包商应确保软件按照本 OPZ 及所提供的文档运行。所提供软件的许可证为永久许可。如使用任何付费许可证或有效期有限的许可证，应在报价中列明随后48个月内这些许可证的成本。3. 承包商应确保可以延长保修与支持，且价格不高于报价中为额外

年所列的价格，延长期限自订单验收之日起至少为48个月。

4. 在保修期内，软件将享有承包商提供的支持，保证

- 报告响应时间不超过 4 小时
- 严重错误的修复时间不超过 48 小时 • 非严重错误的修复时间不超过 10 个工作日 • 此外，每年提供指定的不少于 40 小时的专家咨询，用于故障排除和软件更新。 • 此外，在支持期内，如果订购方扩展网络（从 1GE 扩展至 40GE），承包商将免费进行一次网络重新配置。该重新配置将在双方商定的日期进行，订购方须至少提前 14 天告知。 • 承包商将支持管理员将 OpenStack 版本升级至下一个“扩展维护”版本。
- 承包商承诺支持管理员使用从 Enigma 模块释放的资源扩展云（承包商将与管理员合作，使用新资源（服务器、交换机）扩展现有解决方案）。

### 2.9 实施与付款计划

OPZ 第 1 至 8 点所述功能将在协议签订之日起不超过 180 天内实施。

培训将在协议签订之日起不超过 180 天内，按照培训计划开展。

### 2.10 订单验收方式

订单的执行情况应通过合同附件中规定的验收测试予以确认。

==========


## Acknowledgment

参与这项工作的 ICM 团队成员包括：Marek Michalewicz（项目发起人兼总项目经理）、Wojciech Sylwestrzak（项目管理）、Marzanna Zieleniewska（公共采购）、Jarosław Skomiał（网络与云架构）、Grzegorz Bakalarski、Rafał Maszkowski、Mirosław Nazaruk、Robert Paciorek、Marcin Semeniuk、Sebastian Tymkow（硬件与存储）、Michał Dzikowski、Grzegorz Gruszczyński、Michał Hermanowicz、Maciej Szpindler（软件与云架构）、Joanna Jędraszczyk（用户方面）、Krzysztof Młynarski（网络安全）、Robert Sot（大学流程、规章制度与关系）。这些同事为该项目的技术规范作出了贡献。作者认为，研究和学术生活中的困难与负面经历也应当予以报告，以促进开放、诚实、正直与学习。

作者谨对所有上述同事参与本项目并作出的个人贡献，致以深切的赞赏与谢意。


## References


1. OPENAire 国家倡议——波兰，https://www.openaire.eu/os- poland 2. Armbrust, M., 等（2010）。云计算概览。《ACM 通讯》，53(4)，50-58。3. Buyya, R., 等（2009）。云计算与新兴 IT 平台：愿景、炒作与现实——将计算作为第五种公用事业。《未来一代计算机系统》，25(6)，599-616。4. Gonzalez 等，“云资源管理：面向在复杂基础设施上高效执行大规模科学应用程序和工作流”，《云计算杂志：进展、系统与应用》（2017 13 5 6）：。Beckman, P. H.，“构建 TeraGrid”，（2005）《英国皇家学会哲学汇刊 A 辑：数学、物理与工程科学》363 1602 6 1833 1715 28，101098 2005 2021（）：- DOI：。/rsta. . . 开放科学网格。（）。关于 OSG。检索自 https://www.opensciencegrid.org/，最后访问日期：2023年4月13日 7。XSEDE 美国，htt ps://web.archive.org/web/20220820025648/https://portal.xsede.org/#/guest，最后访问日期：2023年4月16日 8。J. Towns 等，“XSEDE：加速科学发现”，载于《科学与工程计算》，第16卷，第5期，第62-74页，2014年9—10月，80 9 101109 2014 doi：。/MCSE. . . 高级网络基础设施协调生态系统，https://access- ci.org/ 最后访问日期：2023年4月16日 10。PL-Grid——波兰，https://www.plgrid.pl，最后访问日期：2023年4月16日 11。R. Buyya、D. Abramson 和 J. Giddy，“Nimrod/G：全球计算网格中的资源管理与调度系统架构”，第四届亚太地区高性能计算国际会议暨展览会论文集，中国北京，2000，846563 12 283 289 1，101109 2000 页码 - 卷。doi：。/HPC. . . 欧洲开放科学云（EOSC）—— 欧盟委员会发起的一项倡议，旨在构建可信、开放的虚拟环境，以共享和再利用研究数据与服务。（https://eosc.eu/），最后访问日期：2023年4月13日 13。澳大利亚研究数据共享平台（ARDC）——澳大利亚的一项倡议，旨在通过发展开放科学云基础设施，促进研究数据与服务的共享、发布和再利用。（https://ardc.edu.au/），最后访问日期：2023年4月13日 14。国家科学基金会的 CloudBank——美国的一项倡议，为获得 NSF 资助的研究人员提供云计算资源与服务，旨在让研究人员更容易使用云计算。（https://www.cloudbank.org/），最后访问日期：2023年4月13日 15。美国国立卫生研究院的发现、实验与可持续性科学技术研究基础设施（STRIDES）倡议——美国的一项倡议，为研究人员提供基于云的计算和数据资源，以加速生物医学研究。（https://datascience.nih.gov/strides），最后访问日期：2023年4月13日 16。北欧电子基础设施合作组织（NeIC）——北欧国家的一项倡议，旨在为该地区的研究人员开发和运营电子基础设施服务，其中包括云计算资源。（https://neic.no/），最后访问日期：2023年4月13日 17。科学与技术设施委员会（STFC）云门户 —— 英国的一项倡议，为研究人员提供多种云计算资源，包括公有云提供商和私有云环境。https://dareuk.org.uk/，最后访问日期：2023年4月13日 18。Helix Nebula——科学云——欧洲的一项倡议，旨在建立联邦云基础设施，为欧洲

研究人员提供计算资源和数据资源。（https://www.helix-nebula.eu/），最后访问日期：2023年4月13日 19。澳大利亚研究数据共享平台（ARDC）Nectar研究云，https://ardc.edu.au/services/ardc-nectar-research-cloud/，最后访问日期：2023年4月16日 20。Ubercloud https://www.theubercloud.com/，最后访问日期：2023年4月13日 21。https://www.theubercloud.com/ubercloud-compendiums，最后访问日期：2023年4月13日 22。Rescale https://rescale.com/ https://rescale.com/cloud-hpc/，最后访问日期：2023年4月13日 23。XTREME-D https://xtreme-d.net/，最后访问日期：2023年4月13日 24。XTREME Stargate 高性能计算云平台新时代 https://www.youtube.com/watch?v=QdfHnn9L7YY&ab_chann 问日期：2023年4月13日 25。Ronin https://ronin.cloud/，最后访问日期：2023年4月 26。Chameleon：面向大规模边缘到云研究的可配置实验环境，https://www.chameleoncloud.org/ https://scholar.google.com/citations?user=A4pT-T4AAAAJ&hl=en 最后访问日期：2023年4月16日 27。 https://www.bbc.com/historyofthebbc/anniversaries/february/yes- minister/ 最后访问日期：2023年4月16日 28。ACH Cyfronet学术计算中心 https://www.cyfronet.pl/en/4421，2023-29-16 main.html 最后访问日期：4月。智能发展运营计划优先轴详细说明 2014—2020年 https://www.funduszeeuropejskie.gov.pl/media/90780/SZOOP_aktualiza cja 2020 6, czerwiec .pdf 第 . Supercomputing Frontiers 30 Europe 2020年和2021年 https://supercomputingfrontiers.eu/2020/，最后访问日期：2023年4月13日 https://supercomputingfrontiers.eu/2021/conference-programme/，最后访问日期：2023年4月13日 31。NICOLAS TONELLO，“Constellation®——超级计算触手可及——为所有人提供高性能计算能力与专业知识” https://www.youtube.com/watch?v=j1tUJ6PaNf4&t=2s&ab_channel=IC MUniversityofWarsaw，最后访问日期：2023年4月13日 32。K. Kahey，“Chameleon：让科学从云端走向边缘”，https://www.youtube.com/watch?v=ekBAZ1- t9Qk&t=43s&ab_channel=ICMUniversityofWarsaw，最后访问日期：2023年4月13日 33。Wolfgang Gentzsch，“利用分布式高性能计算技术构建自动化、自助式、真正的多云仿真平台”，https://www.youtube.com/watch?v=8uFU59xO6xo&t=18s&ab_channel= 13 2023 ICMUniversityofWarsaw, 34 最后访问日期：4月。Edward Hsu，“采用智能控制平面方法构建面向云的高性能计算” https://www.youtube.com/watch?v=JeOsejMV2mg&t=44s&ab_channel= 13 2023 ICMUniversityofWarsaw, 35 最后访问日期：4月。Tara Madhyastha，“RONIN：云中的安全自助式高性能研究计算”，https://www.youtube.com/watch?v=EmMEd4W0AmU&t=19s&ab_chan nel= 13 2023 36

ICMUniversityofWarsaw，最后访问日期：4月。Valerie E. Polichar，“构想技术愿景：规划新一代研究支持”，https://www.youtube.com/watch?v=osWXOVCNVgU&t=1s&ab_chann el=ICMUniversityofWarsaw 37。数字大学蓝图。加州大学圣迭戈分校研究网络基础设施设计团队报告（2009年4月24日），https://research-it.ucsd.edu/otherresources/UCSD-Blueprint-for-the-Digital-University.pdf 38。Minor, D., & Polichar, V. E.（2022）。蓝图 2030：未来十年的数字大学。加州大学圣迭戈分校：图书馆。检索自 https://escholarship.org/uc/item/920165 2 39 16 页。www.openfoam.com 最后访问日期：2023年4月 40 16。https://www.quantum-inspire/ 最后访问日期：2023年4月

**Marek Michalewicz** Marek Michalewicz于1987年在澳大利亚国立大学堪培拉分校获得理论物理学博士学位。他是新加坡国家超级计算中心的顾问兼技术执行委员会成员，也是开发基于量子隧穿的光电探测器的Siranga初创公司的创始人。他此前还创立并经营过另外四家初创公司。2018年7月至2021年底，他担任华沙大学数学与计算建模跨学科中心主任；2009—2016年担任新加坡科技研究局计算资源中心首席执行官，并在高性能计算与通信组、澳大利亚联邦科学与工业研究组织数学科学部的首席研究科学家（1990—2000年）。他是InfiniCortex项目（2014—2016年）的共同发明者和负责人，该项目旨在建造一台分布于四大洲七个国家、通过InfiniBand（远程直接内存访问）连接的全球规模并发超级计算机。他也是推动在新加坡建立国家超级计算中心的主要架构师和关键人物之一。Marek的研究领域包括超级计算、计算科学、理论物理学和纳机电系统传感器。